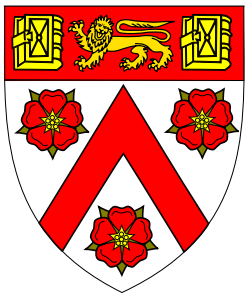
Breaking the Grant Cycle: On the Rational Allocation of Public Resources to Scientific Research Projects

Shahar Avin



Trinity College

This dissertation is submitted to the University of Cambridge
for the degree of Doctor of Philosophy
August 2014

Contents

Declaration and statement of length

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. The dissertation is 79,913 words long and does not exceed the word limit (80,000) set by the History and Philosophy of Science Department.

Abstract

The thesis presents a reformative criticism of science funding by peer review. The criticism is based on epistemological scepticism, regarding the ability of scientific peers, or any other agent, to have access to sufficient information regarding the potential of proposed projects at the time of funding. The scepticism is based on the complexity of factors contributing to the merit of scientific projects, and the rate at which the parameters of this complex system change their values. By constructing models of different science funding mechanisms, a construction supported by historical evidence, computational simulations show that in a significant subset of cases it would be better to select research projects by a lottery mechanism than by selection based on peer review. This last result is used to create a template for an alternative funding mechanism that combines the merits of peer review with the benefits of random allocation, while noting that this alternative is not so far removed from current practice as may first appear.

Acknowledgements

"We are here to help each other get through this thing, whatever it is."
– Mark Vonnegut.

This thesis would never have reached the submission stage without hard work and sharp intellect from my supervisor, Tim Lewens, and my advisor, Stephen John. They have been very patient, yet strict, as I fumbled around blind in the first two years of my PhD in a topic that was far too big and ill-defined. Once I found my focus they helped me maintain an accelerated pace, regularly providing insightful comments and guidance through wave after wave of countless drafts. Any imperfections left in the text now in your hands are entirely my fault, and are present in spite of the best efforts of these two remarkable gentlemen.

I would like to thank the Master and Fellows of Trinity College for their willingness to fund three years of doctoral research on the topic of research funding. I would especially like to thank Rosemary Jolley for all her help, and the accommodation office for extraordinary assistance in a time of need. Eight years in Trinity are a very long time, still I hope I will be back one day.

My shift from physics to philosophy of science was far from expected, and I would like to thank four individuals in particular for being instrumental in this transition: Sachiko Kusukawa, Trinity's HPS director of studies; Peter Lipton, my first philosophy of science lecturer, who tragically passed away that very year; Mike Collins, my IB philosophy of science supervisor; and Marina Frasca-Spada, who was willing to supervise a Part II physicist's Part II HPS dissertation. As Sachiko would say, I "caught the HPS bug" fairly early on, though it took me some time to realise that.

I am very lucky to have discovered Cambridge HPS. I was a member of the first HPS Part III cohort, and I am grateful to those who made that institution a reality. The Part III/MPhil cohort of that year was an inspiring bunch, and I throughly enjoyed our regular pub outings and the many friendships that emerged. I would especially like to mention Becky, Dima, Eoin, Jesse, Katharina, Kathryn, Lavinia, Mike, Nick, and Petter. It was a good year.

The year following proved more challenging. It was, however, a remarkable opportunity for people to show their caring side, and I'm indebted to Amanda, Dan, Dima, Katharina, Illan, Netta, Vivek, and Walter, as well as the Israel crowd I will mention later. On a more upbeat note, it was the year I started working as Hasok Chang's research assistant, and the year I started supervising, both experiences which have been extremely rewarding. I am very fortunate to have met Hasok, to whom I am grateful for introducing me to exciting avenues of research, and for showing such a nurturing attitude towards my work and my growth as a philosopher. I hope I managed to reflect the model set out by Tim, Stephen, and Hasok towards my own students in supervisions over the years.

Through Hasok I got to know the Society for Philosophy of Science in Practice, a truly wonderful group of passionate academics. It has been a pleasure working with the committee on the society's website, and the two SPSP conferences I attended in Exeter and Toronto have had a strong impact on my view of the world, as well as being very good fun. Toronto was also the first time I got to present my work to an international audience, and I'm grateful to Maureen O'Malley, Chris Haufe, and Greg Petsko for organising the symposium on science funding, and to the audience for very helpful comments. I'm also grateful to Maureen for her encouragement and support that led to my single publication to date. I would also like to express thanks for helpful comments from the audiences at two Philosophy Workshops in Cambridge HPS and a talk at an EPSA conference in Helsinki.

Outside of conferences and workshops, I had great discussions and received very helpful comments on my work from busy and talented people, and would like to thank Anna Alexandrova, Tiago Mata, Michael Strevens, Eyal Maori, Aviv Bergman, Jeremy Butterfield, Hasok Chang, Stuart Firestein, Craig Bakker, Alistair Potts, Nadanai Laohakunakorn, Vivek Thacker, Yuval Kantor, Netta Chachamu, Vashka dos Remedios, Minwoo Seo, Eoin Phillips, Dmitriy Myelnikov, Katharina Kraus, and Fiona Blower. I'm sure there are names that are missing from this list, so if I ever pestered you with my views on science funding (and I probably have), consider your name included above.

Doctoral studies and long winters can be a recipe for emotional difficulties. My gratitude, and apology, is offered first and foremost to Fiona, who heroically suffered my presence every day for a significant portion of these four years, and made me a better person in the process. Further thanks are due to all those who listened to my silly complaints, brought a smile to my face, and in general helped me get through this thing, whatever it was. Take a deep breath, because this will be long (and still very partial). The Mond crowd: Vashka, Minwoo, Eoin, and Katharina. The Physics crowd: Vivek, Nads, and Nils. Other Cambridge people: Zoya, Han-Ley, Dima, Emily, Eyal, and Keren. London people: Maor and Dan. The Gym crowd: Mila and Shahar, Roy and Moran, Netta and Eran, Yaara, Noam, and Oren. Spinning and Tripping: Nadav, Nadav, Yoav, and David. Other army people: Yuval and Mari, Sivan, Shai, Omri, Neetai and Daphna, Tomer and Tamar. School people: Shai, Liran, Nitsan, and Elad.

Finally, a massive thanks to my family, a collection of warm, loving, wonderful people: Zeev, Sagit, Dori, Yahli and Ori, Erez, Natali, Opal and Orian, Yafa, Tamar, Gal, Reut and Maayan, Rotem, Ayelet, Tuli, Mary, and Nuli. In loving memory of Aunt Kalya, Grandpa Max, Grandma Lala, Grandpa Aryeh, and Grandma Hinda. And to conclude, a big thanks to my parents, Sarah and Yankale, who brought me *ad halom*.

Introduction

In the spring of 2013 the internet was flooded with comments over a $384,949 grant awarded by the United States' government for the study of duck genitalia (Brennan, 2013). The money was allocated by the National Science Foundation, a US federal agency. The decision to fund the project was made using grant peer-review, a process that selects scientific projects for support based on an evaluation of the project proposal, conducted by scientific peers of the project's principal investigator. As expected, the interest in the story was short-lived, and the issue of public support of basic research once again returned to receiving little interest from the general public. Perhaps more surprisingly, in the philosophy of science community there is a similar baseline lack-of-interest in science funding decision-making. This lack of interest is surprising for two reasons: first, it is surprising because scientific outputs, which rely on funding, have overall an important influence on quality of life in contemporary societies; second, it is surprising because of the central role that funding decisions play in the lives of practising scientists, due to the time and emotions invested and the crucial role they play in scientific careers.

This thesis breaks away from this baseline by providing a detailed philosophical study of science-funding decision making. The thesis addresses the following question:

Are the processes used by public science funding bodies to make funding decisions rational, and can they be made more rational?

To address this question, the thesis starts with an evaluation of current grant peer review practices. While accepting the claim that, from all persons in the society, scientific peers are in the best possible position to evaluate scientific merit, I also appreciate the strength of an existing criticism of peer review based on poor reliability and bias in the evaluation of merit. To reconcile these conflicting positions the thesis develops a more comprehensive account of scientific merit, first by considering existing models of project choice and then by developing a novel model based on a causal notion of "epistemic fitness". Using this notion, the thesis advances a sceptical argument regarding the predictability of scientific merit, given the highly complex and dynamic nature of epistemic fitness. The sceptical argument is further refined using computer simulations of different funding mechanisms and their effect on the accumulation of epistemic fitness over time. Based on the sceptical argument and the results of the simulations, I propose an alternative to grant peer-review, a mechanism that combines peer evaluation with elements of random selection.

Chapter 1 discusses science funding by peer review. The chapter is composed of three parts. The first part presents an overview of peer review in the wider scene of Research and Development, the key features of peer review, and the historical origins of the current system. The second part builds on existing survey work to derive a set of desiderata for a science funding mechanism, based on decades of experience with peer review. The third part presents the most sophisticated

philosophical defence of science funding by peer review (Polanyi, 1962), and an important recent philosophical critique of the practice (Gillies, 2007, 2008, 2009, 2012, 2014). Like this thesis, Gillies is sceptical of scientists' ability to correctly judge their peers' work. Gillies' work focuses on biases in peers' evaluation of merit, but does not provide a detailed account of what merit is. This limitation is addressed by the work presented in this thesis, which presents a detailed argument that supports, to an extent, Gillies' proposal of funding science by random selection.

Chapter 2 discusses the existing literature in philosophy of science dealing with models of scientific project choice. The chapter starts with a defence of the choice of a model-based approach for the task of evaluating science funding decisions. The chapter then surveys the history of models of science project choice, divided into two "generations". The first generation of works include Peirce (1879/1967); Kitcher (1990, 1993) and Strevens (2003), and shows strong similarity to rational-choice models. The second generation is exemplified in the work of Weisberg and Muldoon (2009), moving away from simulating single-agent decisions to a simulation of a population of investigators on an epistemic landscape. The works are evaluated for their applicability to the science funding case, noting improvements from the first generation to the second in the reduction of the number and scope of unrealistic assumptions. Nonetheless, Weisberg and Muldoon's model is deemed inappropriate for direct application to the science funding case, for two main reasons, both relating to their notion of "epistemic significance" which measures the merit of potential research projects. The first problem is that epistemic significance is not related to any notion of contribution to well-being, which is important when evaluating decisions made by public funding bodies. The second problem is that epistemic significance is assumed to be fixed over time. These two problems are addressed, respectively, in the two chapters following.

Chapter 3 constructs novel models for representing the impact of research. A stepwise model-building strategy is presented, which first constructs the "information landscape" model as a representation of the value of different corpuses of information. The construction of this model is informed by a detailed study of "fitness landscape" models in population biology, and by a study of a new (non-evolutionary) concept of scientific merit, labeled "epistemic fitness". The information landscape model is then used to construct a revised version of Weisberg and Muldoon's "epistemic landscape" mode. The chapter concludes with a demonstration of how the revised epistemic landscape model can be used to assess funding decisions within their social context.

Chapter 4 argues for the inclusion of time-variation of epistemic fitness in models of funding decisions. The chapter starts with a brief exploration of the philosophical connection between ignorance and decision making. Then, using three historical examples and the models developed in Chapter 3, the chapter presents a list of processes by which epistemic fitness can change over time, as a response to ongoing research activity. Since *ex ante* decisions about the relative potential of different research projects are essentially predictions, the presence of dynamic processes provides sources of uncertainty that can undermine the accuracy of these predictions.

Chapter 5 presents a computer simulation of different funding selection mechanisms. For the purpose of simulation, a revised version of Weisberg and Muldoon's epistemic landscape model is used, updated to include selection by a central funding agency and processes of fitness dynamics. The simulated funding mechanisms include an optimal "god's-eye" selection process, a limited-vision optimisation selection (representing selection by peer review), and random selection

by lottery. The results show that on small and static landscapes limited-vision optimisation outperforms a lottery, but on large dynamic landscapes a lottery outperforms limited-vision optimisation. The chapter concludes with a discussion of the idealisations in the simulation, the resulting limitations, and potential future improvements.

Chapter 6 presents a template for an alternative science funding mechanism that combines peer review with elements of random selection. The chapter is composed of three parts. The first part, building on the results of the computer simulation, reinforces the criticism of peer review using recent empirical studies of the (limited) reliability and (high) cost of peer review (Graves et al., 2011; Herbert et al., 2013). The second part presents a theoretical background on the use of lotteries, both as fair and effective selection mechanisms (Boyle, 1998) and as economic distribution mechanisms (Boyce, 1994). In the third part, the empirical evidence and theoretical background are used to construct a template for a novel triage mechanism. In this mechanism, expert assessment is used to triage short proposals into rough categories: high/medium/low fitness, or unknown fitness. Proposals of high fitness are funded, and proposals of medium or unknown fitness are admitted into a lottery. The chapter concludes with a consideration of some of the limitations of the lottery proposal, while showing that in many cases where we would expect a lottery to perform poorly so would peer review.

Chapter 1

Science Funding by Peer Review

No-one can assess
proposals better than peers,
but are they biased?

As outlined in the introduction, this thesis is concerned with the evaluation of the processes used to make decisions about the allocation of resources to research. In any evaluation, we require some criterion by which to evaluate; in the case of funding mechanisms, we need to explicate the desiderata that should be met by these mechanisms. In general, there are (at least) two ways to elucidate these desiderata: we could try to come up with an *a priori* list of desiderata, or we could look at an existing implementation of a funding mechanism, and by studying its operations and its consequences infer general desiderata for funding allocation. In this chapter the latter route is followed. By examining the operation of science funding by peer review, we arrive at a set of desiderata for funding allocation mechanisms. These desiderata are then used to reflect back on the peer review mechanism, setting up a concern about its optimality that will provide the main thrust of argument throughout the thesis.

The chapter is composed of three parts: the first part presents the history and current implementations of peer review, against the wider context of resource allocation for research and development (R&D); the second part discusses the desiderata for a science funding mechanism, in light of the accumulated experience of scientists and politicians with the peer review system; the third part reflects on whether peer review meets these desiderata, by discussing Polanyi's defence of peer review, and Gillies' criticism of this mechanism.

1.1 Contemporary science funding

Investment in Research and Development (R&D) is a substantial global phenomenon. In developed countries, such as the United States of America (USA), Japan, South Korea, and the western member countries of the European Union (EU), spending on R&D is often in the range of 2-3% of the Gross Domestic Product (GDP).[1] This number takes into account both public (government) spending and private (industry) spending. When taking the significant GDPs of these countries into account, we arrive at a global R&D investment of many hundreds of billions of dollars. The sheer magnitude of R&D expenditure, along with unique features of the nature

[1] R&D statistics in this section are from Kennedy (2012), and represent data for fiscal year 2009.

of scientific research, makes the study of the strategies used to invest funds within this area a worthwhile project.

Spending on R&D is not homogeneous: it can be divided by source of funding, and by type of research. While boundaries between divisions may be blurry, certain distinct categories of R&D funding emerge. The rationale for funding in each category, and the philosophical analysis of desiderata and appropriate mechanisms, may differ between these categories. This section makes a quick survey of the key categories of R&D spending, as they are defined and acted upon from the perspective of R&D policy research. To make the presentation more tangible it is accompanied by recent statistics regarding R&D in the USA. The USA was chosen both for availability of data[2], and because the USA is a world leader in R&D expenditure.[3] The section concludes with an explanation of the decision to focus in this thesis on one particular category of R&D, that of publicly-supported basic science.

1.1.1 Public versus private spending on R&D

The first important distinction to be made within R&D spending is between the sources of funds: public funds, generally originating from tax collection and allocated by the government, and private funds, generated mainly from corporate profit or charitable donations, and expended by for-profit or not-for-profit private organisations. The distinction between public and private expenditure on R&D is significant for several reasons. When evaluating the effectiveness of an R&D funding strategy within an institution, it is important to consider whose interests should be addressed by this type of institution. While a for-profit firm exists to address the interests of its owners, shareholders, and potentially its employees (within regulatory bounds), and a charity exists to address the interests of its managers, donors, and beneficiaries (again, within regulatory bounds), a government should address the interests of all citizens of the country.[4] This consideration of the interests and needs of the entire citizenry makes the evaluation of individual research projects harder and less clear-cut; in the public funding case we are much more likely to run into clashes of incommensurable values than in the private funding case, simply because the number of preferences that need to be balanced increases as the number of individuals involved increases. Furthermore, the common ground for shared values is often greater in the private sector, e.g. around the company's or charity's mission statement.

In contemporary USA R&D expenditure, the private sector accounts for about two thirds, with the dominant provider by far being industry, whereas the public sector funds one third of R&D.[5] Thus, about twice as many funds in the world's largest R&D nation are determined according to the interests of shareholders and company owners, which may or may not be aligned with the interests of the population as a whole. Most of these funds are not directed towards what we often consider as "science", but towards what we would consider "technology", as discussed

[2]The National Science Foundation in the USA collects and regularly publishes extensive statistics on R&D expenditure in the USA and across the globe.

[3]Total national R&D expenditure: USA $401.6 billion (largest), China $154.1 billion (second largest); All EU nations combined $297.9 billion.

[4]This is assuming, of course, we are dealing with a democratic or socialist government. The thesis does not consider governments which are concerned only with the interests of a small minority of the population, such as despotic or oligarch governments. In addition, this picture also ignores the many complexities arising in practice where firms, charities or governments address other interests from the ones listed above, or when their actions are biased towards the interests of particular groups.

[5]USA R&D expenditure: industry 62%, federal government 31%, private non-profit organisations (including private universities) 6%, state and local government less than 1%.

below.

1.1.2 Basic versus applied science versus development

A second distinction within R&D expenditure is the kind of research or development work taking place. While assignment of individual projects into any category can prove challenging, most within the R&D policy world recognise categories that are similar to the three characters of work defined by the USA National Science Foundation (NSF):

> **Basic Research** Research that seeks to gain more complete knowledge or understanding of the fundamental aspects of phenomena and of observable facts, without specific applications toward processes or products in mind.
>
> **Applied Research** Research aimed at knowledge necessary for determining the means by which a recognised need may be met.
>
> **Development** The systematic use of the knowledge or understanding gained from research, directed toward the production of useful materials, devices, systems, or methods, including design and development of prototypes and processes.
>
> (Kennedy, 2012, pp. 4-5, changed styling for clarity)

It is clear from these definitions that basic research is closer to what we commonly consider "science", and development closer to "technology", with applied research occupying a fuzzy middle ground. Not much rests on these labels, and the search for a precise definition of these terms, and the boundaries between them, lies well outside the scope of this thesis. However, they can be useful, in as much as this thesis relies on an existing body of scholarship in the philosophy of science, which can only be applied straightforwardly to a subset of R&D projects. This subset significantly overlaps with projects which are commonly categorised as basic research.[6]

Basic research, to which this thesis applies most directly (see below), is only a minor component of global R&D. As discussed above, about two thirds of R&D funds in the USA are provided by private industry. Almost all of these funds are directed towards technological development in few high-tech sectors, and very little is directed towards basic research.[7] The major supporter of basic research in the USA is the public, via the federal government. However, even the government does not direct a majority of its funds to basic research, as its R&D portfolio contains a significant development component, for industries that cannot easily be privatised, such as defence and energy.[8] Nonetheless, the largest institutions supporting basic research, both in the USA and in the world, are federally funded agencies, of which two are particularly dominant: the National Institutes of Health (NIH), and the NSF.[9]

1.1.3 Why focus on publicly-supported basic research?

This thesis is mostly concerned with decisions concerning the allocation of public resources for basic research, e.g. the decisions made by NIH and NSF. The focus on this category of

[6] For an overview of the distinction between philosophy of science and philosophy of technology see Franssen et al. (2013). While there is a debate whether technology is the same as applied science (Bunge, 1974), or something distinct from it (Simon, 1996), there seems to be general agreement that basic scientific research is a different kind of activity from technology development, and requires different philosophical treatment.

[7] USA industry R&D: 80% development, 14% applied research, 6% basic research.

[8] USA federal R&D: 52% development, 25% basic research, 23% applied research.

[9] Expenditure on basic research: NIH $18.9 billion, NSF $5.6 billion.

R&D spending follows from the kind of argument presented, a sceptical one. Over the following chapters, it will be argued that science funders do not have sufficient information to make justifiable decisions regarding resource allocation. The extreme case for this argument is the case of publicly-supported basic research, as it offers the broadest combination of sources of uncertainty: the very wide spectrum of values being pursued on behalf of the public, and the long delay between research and payoff characteristic of basic research. In a sense, this presents a weak form of the argument, as it is not claimed that the argument extends to cases of less uncertainty, such as applied research or privately-funded research. However, even this weak argument, if accepted, can have significant effect on real life practices. In addition, while they are not discussed directly, consequences of the argument can be envisaged for other, less uncertainty-ridden categories of R&D spending.

1.2 The process of funding by peer review

The previous section showed that the leading institutions in funding USA basic research, and global basic research, are the NIH and the NSF. Both agencies allocate funding by a scheme of project choice called *peer review*, where research proposals originating from practising scientists are reviewed and ranked by their peers. The following summary of the operation of peer review is based on consideration of the application and review process at several governmental science funding agencies, including NIH (2013a,b), NSF (2013a,b), the Australian National Health and Medical Research Council (Graves et al., 2011), and the Austrian Science Fund (Dinges, 2005).

The process of resource allocation for basic research by peer review is the dominant contemporary form of resource allocation for scientific projects, the "gold standard" of science funding. Some aspects of the process are strongly conserved across nations and institutions:

Investigator freedom Project proposals originate from the investigators, not dictated by the funding body or a central organising committee. The extent to which investigators are free to design projects is limited under various guideline constraints, but there are many opportunities for significant levels of freedom.

Individual projects As proposals originate from the investigators, they arrive at the funding body as discrete, compartmentalised funding opportunities. The funding bodies have the role of choosing among them, but they do not, to any significant extent, coordinate between different investigators to form overarching research programmes.[10]

Information provision As proposals originate from the investigators, they must inform the funding body about the contents and merits of their proposed projects. This is often done using a detailed written research plan, accompanied by various supporting documents.

Peer assessment Funding bodies seek the expert opinion of one or more scientists in evaluating the merit of the proposed projects. While there are guidelines for component categories of evaluation, the decisions are still significantly subjective, not algorithmic or box-ticking.

[10]From conversations with practising scientists it seems that funding bodies now encourage, and expect, investigators from different fields to make joint proposals, to indicate "synergies" that make the research team more "dynamic". However, these connections between scientists are made by the scientists themselves, and occur prior to the submission of the proposal. Once the proposal is submitted the funding bodies do not actively make connections between the authors of different proposals.

Integration of assessments Often assessment is sought from more than one source, e.g. multiple reviewers or a mix of internal and external review. The different assessments are always combined in some way to form a single judgement per proposal, which is then compared to the judgements of other proposals.

Ranking and cutoff There are never enough resources to fund all projects proposed. As such, comparisons of integrated assessments are used to decide which projects will get funded and which will not.

Other aspects of the process exhibit more variability, such as the identities of reviewers, the method of integrating assessments, and the guidelines for merit evaluation. The focus of this thesis, however, is on the common features of peer review, and the rationality of making decisions about funding using this process.[11] The practice of science funding by peer review is very strongly entrenched, as it has been around for a long time, and so opening it up for examination may seem strange. However, from a historical perspective, the contemporary process of grant peer review, and its control of very large funds, is a relatively recent phenomenon, originating during the Second World War (WW2). The next section presents the story of the historical origins of USA governmental support for basic science by peer review.

1.3 Historical origins of the current system

The numbers show that contemporary science funding is, dominantly, American science funding (§1.1). This is even truer if we consider that many countries have adopted, with various modifications, the American model of support for science at various stages following WW2 and the establishment of the American mechanisms for support of basic research. This section presents a historical account of how these American institutions came into existence in the period during and shortly after WW2. The account is adapted from two comprehensive books on the politics of American science by long-time science journalist Greenberg (1999, 2003).[12]

1.3.1 Before and during WW2

In the pre-war era, the American congress would occasionally support applied research for specific goals with an expected short-term payoff, but rarely provided support for basic research, or the acquisition of knowledge for knowledge's sake. Sources of funding for basic research were mainly philanthropic, including the Smithsonian Institution, founded with the wealth of a British scientist for the "increase and diffusion of knowledge among men", the Rockefeller Foundation, financed through the wealth of the Rockefeller family (mainly from the oil industry) to support science and health research and medical education, and the Carnegie Corporation, founded by a railroad industrialist for "the advancement and diffusion of knowledge and understanding".

Early public funding bodies for basic research were formed after WW1, including the National Advisory Committee for Aeronautics (NACA) and the National Cancer Institute (NCI). However, both philanthropic and public research budgets pre-WW2 were only a fraction of the funds available for research post-WW2, even when corrected for inflation (Greenberg, 1999, p. 59).

[11] The meaning of rationality used in the context of this thesis is discussed in §1.4 below.

[12] Greenberg presents a clear and accessible account of the relevant period, at the right level of detail for this thesis. A more detailed account, embedded in the historical analysis of the period, is given by Agar (2012, especially pp. 264-6 and pp. 302-8).

During WW2, the aspirations of the scientific community to join the war effort, combined with Roosevelt's anti-isolationist desires, led to the foundation of the National Defence Research Committee (NRDC) which a year later transformed into the Office of Scientific Research and Development (OSRD). This was a significant federally funded research organisation that employed civilian scientists for military-relevant research.

While the familiar story is of Szilard's and Einstein's letter to the president,[13] the real driving force behind the establishment of OSRD was Vannevar Bush, then president of the Carnegie Institution in Washington, and the alliance he formed with other leading figures of institutional science: James B. Conant, president of Harvard, Karl Compton, president of MIT, and Frank B. Jewett, president of the National Academy of Science (NAS) and of Bell Telephone Laboratories. Bush was made director of NRDC, and then OSRD, following its formation.

Several significant precedents were set by OSRD:

- Military-relevant research was conducted by civilians, who were not taking direct orders from the military.
- Significant federal funds were under the discretion of career scientist-administrators.
- A contract system was established for supporting civilian academic research using federal funds, which did not restrict the academic freedom of the recipient of the funds: the contracts specified the amount of funds, the general problem, and a date by which a report would be submitted, but not the methods to be taken to explore the problem.

These contracts, designed by Bush in OSRD, managed to strike a successful balance between the academics' desire for intellectual freedom and the government's desire for accountability.

Musings about the role of government support for post-war science began in earnest around 1944, with the winding down of research activity in some of the war-established research laboratories. Motivated by early proposals from Washington, the politicians of science in Washington[14] arranged for president Roosevelt to ask Bush to draft a letter, signed in advance from Roosevelt to Bush, asking Bush to raise the key issues of post-war research. In the letter, dated November 17, 1944, Bush proceeded to ask himself, under the auspices of the president, to inquire into how "the information, the techniques, and the research experience of the Office of Scientific Research and Development and by the thousands of scientists in the universities and in private industry, should be used in the days of peace ahead for the improvement of the national health, the creation of new enterprises bringing new jobs, and the betterment of the national standard of living" (Bush, 1945, p. iii).

Bush's reply was a synthesis of panel reports, whose membership was largely drawn from OSRD staff. The reply, presented to president Truman, Roosevelt's successor, on July 5, 1945, was the famous report *Science, The Endless Frontier* (Bush, 1945). In the report, Bush presented two justifications for government support of basic research, one consequentialist, the other deontological. On the consequentialist line, Bush extolled, with examples, the contributions of basic science to advancements in fighting disease, obtaining military superiority, gaining an

[13] Convinced by Szilard, Einstein wrote a letter to Roosevelt to inform him of the potential, and potential risk, of nuclear weapons. While raising Washington's awareness of the issue, the direct result was a poorly-funded fact-finding mission.

[14] Probably either Compton, mentioned above, or Albert Lasker, a businessman and future benefactor of the NIH, and his wife Mary Lasker, a key figure in medical politics.

economic advantage, and creating jobs. Bush claimed basic science was *necessary* for these advancements, as they rely on utilising novel knowledge, and therefore the extant body of knowledge must be continually rejuvenated (via basic science) to ensure sustained improvement. Such advancements are clearly in the interest of the public, and therefore in the absence of alternative sources of funding for basic science, the government should step in. Bush concludes the consequentialist line by explaining why alternative sources of funding are unlikely to emerge: funding from other countries is unlikely following the ravages of war in Europe, and industry funding is unlikely due to the high risk and long delays in payoff involved in basic research. On the deontological line, Bush claimed that public support of science is an extension of the federal government's commitment and obligation to support the exploration of new frontiers, with the frontiers of knowledge taking the place of the (largely-explored) physical frontiers in land and sea.

In relation to Bush's report, this is a good place to state the position of this thesis regarding the benefits of science to society, and the role of the government in supporting basic science. In this thesis a weak form of Bush's argument is accepted without argument: that science *does* contribute to social well-being more than it costs (though not infallibly), and that it is beneficial for the public to support basic research. The next chapter clarifies why, in the context of this thesis, this can be accepted without argument, as the thesis is concerned with how to make specific decisions about allocation to projects *once a general decision about support for science has been made*. Furthermore, Chapter 3 shows ways in which the tools developed in this thesis can be used by someone holding a different set of values, including about the historical and potential benefit of science to society.

1.3.2 Post WW2

Despite Bush's attempts, the National Science Foundation was not established until 1950, and even then, it was initially given only a small budget. The reasons for the delay focused on disagreements about the level of accountability to government, the identities of key positions in the governance of the foundation, and an explosion in the political debate about the control of nuclear energy following the nuclear bombing of Hiroshima and Nagasaki.

In the years directly following the war, Bush's proposal for a National Research Foundation received cold responses from Truman's White House, and Bush himself became more marginal in the Washington scene. Truman appointed John R. Steelman to draft a new proposal for the federal support of research, and it was Steelman's report, *Science and Public Policy*, delivered to Truman in 1947, which largely shaped the final structure of NSF. Nonetheless, general themes, such as investigator independence and government support for basic science to promote economic growth, were maintained.

Over the following years and decades, both new federal agencies, such as the Atomic Energy Commission (AEC), the Department of Energy (DOE), and the National Aeronautics and Space Administration (NASA), and old ones, such as the Department of Defence (DOD), and the National Institutes of Health (NIH) expanded their R&D budgets, including an expansion of their spending on basic research. Meanwhile, NSF budgets have also grown. The growth in federal R&D expenditure, as well as the economic growth of the USA, has been so significant that the USA is now the world leader in both total R&D spending and in basic research spending, as discussed in §1.1.

This concludes the exposition on the state and origin of peer review, the leading contemporary mechanism for public support of basic science. The next section deals with the lessons that have been learned about the operations of peer review, and the desiderata for a science funding mechanism in general, from several decades of experience with peer review.

1.4 Desiderata for a public science funding mechanism

The aim of this thesis is to evaluate the rationality of science funding decisions. In the context of this thesis, a rational decision is defined as one which the agent has good reason to believe will bring her closer to the goal associated with the action.[15] Since funding decisions are made continuously, it is more interesting to inquire about the rationality of the strategy used to make these decisions, e.g. by relying on peer review, than the specific rationality of individual funding decisions. Given two possible strategies for allocating funds S_1 and S_2, and some aim A, the funding body would be rational to adopt S_1 over S_2 if it can present an argument that provides good reason to believe S_1 is more likely to bring about A than S_2. If the aim is a general evaluation function, e.g. the generation of useful scientific results, it may be the case that both S_1 and S_2 lead to a state that realises to some extent the aim A (i.e. both lead to the generation of useful scientific results); in such a case a rational adoption of S_1 will require an argument that provides good reason to believe the amount of A generated by S_1 will be larger than the amount of A generated by S_2. Thus, any evaluation of the rationality of a funding body's adopted strategy would require an establishment of the goal the funding body is attempting to pursue.

This chapter presents two arguments regarding the rationality of making funding decisions by peer review, and later chapters develop a third argument. These different arguments assume slightly different aims for public science funding bodies.[16] However, in the various arguments, appeals are consistently made to a group of desiderata relating to the performance of a public funding mechanism. While substantive, these desiderata are also vague enough to accommodate the slightly different aims associated with each of the arguments. These desiderata largely overlap with the desiderata presented by Chubin and Hackett (1990), as described below.

Chubin and Hackett (1990) present a critical examination of peer review, based on evidence from surveys of practising scientists. Their stated aim is to overcome the nearly-mythical standing of peer review as a pillar of modern science, and highlight the fact that very little has been done to subject peer review to methodical analysis, despite known tensions and probable shortcomings. The key findings of concern to this thesis are the ones regarding grant peer review, or the use of peer review for allocation of research funds. Chubin and Hackett consider specifically grant peer review in the two organisations that were highlighted in §1.1: NIH and NSF. The findings of interest have been summarised in a paper by Chubin (1994).[17]

[15] For the epistemologist, this can be restated in more familiar terms: a rational action is one for which the agent has justification for the action's reliability in producing the associated outcome. For an elaboration of this point, and discussion of some of its subtleties, see Davidson (1980).

[16] The argument developed in this thesis assumes that the aim of public funding bodies is to increase social well-being via the scientific generation of new, reliable information. An elaboration of this assumption and the use of terms is presented in §3.1.1.

[17] In the period between 1990 and 1994 Chubin moved from a Congressional office, where he was in charge of reviewing grant allocation, to an internal review unit of the NSF. Unsurprisingly, the tone of the 1994 paper is more mellow than the tone of the 1990 book. Chubin and Hackett have themselves commented on this phenomenon when comparing studies conducted within and without the funding agencies. However, we are interested in the content of Chubin's arguments, not their tone, and these are consistent across the two documents.

The main argument of Chubin and Hackett is that peer review serves a function for multiple stakeholders, each having slightly different expectations from the process and its products. These different desiderata are often in tension with each other, and so the process of peer review often fails to fully satisfy any of the desiderata, to the chagrin of stakeholders. Chubin and Hackett note, using survey data, an increase in the concern scientists and other stakeholders report regarding peer review. However, they note that as total success rates in grant peer review decline, because the increase in the scientific cohort size outpaced growth of allocated funds, pressure increased within the scientific community, which led to increased scrutiny of the allocation mechanism, though it was not any feature of peer review *per se* that caused the increase in pressure. Nonetheless, the increased attention to peer review brought to the fore explicit statements about what different parties considered the proper function of peer review, and what were the perceived shortcomings in fulfilling this function.

Chubin presents a summary of seven different desiderata that apply to mechanisms that distribute public funds to researchers:

Effectiveness The mechanism should be effective at identifying high quality research. This is clearly a primary desideratum. The results of effective allocation would be that high quality research is supported, leading to scientific progress and new knowledge. Chubin points out that a sound study of effectiveness would require an *ex post* evaluation of both funded and unfunded proposals, but that funding agencies are reluctant to perform such studies.[18]

Efficiency All parties involved would prefer, *ceteris paribus*, for the process to be as efficient as possible, meaning for it to require as little time and resources while providing similar levels of effectiveness. This applies both to the time and resources on the reviewing side (administration, internal and external reviewers) and the applicants' side. Chubin notes an increasing concern regarding efficiency, especially in light of falling success rates, as applicants are less willing to invest a significant amount of time and emotional energy into applications that have a base-line success rate of 25-40%.[19] In addition, worries regarding efficiency in peer review also increase as other, unrelated, bureaucratic hurdles increase, e.g. forms and regulations regarding human and animal subjects.

Accountability The taxpayers and their representatives want the allocation process to maintain accountability, to make sure scientists are spending the funds in ways that would further scientific research, to make sure they remain within the bounds of the project outlined in their proposal, that due process is followed in the allocation process, and that laws and regulations are adhered to, e.g. regarding treatment of human and animal subjects. Scientists are often wary about these requirements, raising concerns of red tape and restrictions preventing them from pursuing unpredictable promising leads part-way through a grant. The resulting mechanisms, e.g. in NIH and NSF, aim to strike a balance that on the one hand entrusts operational responsibility of individual project choices to reviewers, but on the other hand gives government officials overall responsibility for the research portfolio, and these officials are accountable to Congress.

[18] An example of institutional objection to studying failed results is the refusal to release names and contact details of authors of rejected proposals for survey purposes. A legal episode involving Chubin, that eventually led to the disclosure of these details, is recounted in the book.

[19] Since the 1990s success rates have dropped further; see the discussion of recent empirical studies in Chapter 6.

Responsiveness Governments and practising scientists, especially younger ones, want the allocation mechanism to be able to respond quickly to new challenges or new promising avenues of research. They worry that the process may entrench existing disciplines and direct funds towards established researchers, either through risk aversion at the management level of funding agencies, or through an "old boys network" of established researchers that bias reviews in favour of their contacts and against newcomers. On the flip side, established researchers are concerned with the emergence of fads and fashions, directing money away from established research programs towards new fast-growth disciplines. Chubin notes that there is often little agreement about what is a fad or fashion, or what is a stagnating field that is being sustained by overly conservative investment.[20]

Rationality [21] The allocation process should be understood by its participants, and should be seen as rational, or justifiably likely to bring about its purported aims; in other words, its rationale should be transparent to those taking part. If the workings of the allocation mechanism seem mystical, secretive, or unfathomable, then the willingness of scientists to cooperate with it, to tolerate rejections, and to invest their time in reviewing, will decline. Chubin notes that funding agencies have responded to criticism of lack of apparent rationality, and have started to present more information about the process to the scientific community, including explicit review criteria and feedback to applicants.[22]

Fairness The allocation mechanism should distribute funds fairly, meaning it should not unjustifiably discriminate against any individual or group. While some consider this to mean evaluation should be based only on the merit of the project, others consider fairness to mean a deliberate drive towards greater diversity and representation of minorities in the funded cohort, even at the expense of apparent quality. Chubin notes the emergence of some "set aside" programs in NSF, where only applicants from a certain background may apply, and that these programs have attracted much criticism. The issue of fairness in resource allocation is discussed further in Chapter 6.

Reliability The allocation mechanism assigns a comparable measure of scientific merit to proposed projects, and can therefore be considered to include a measurement component. This measurement component of the mechanism should meet general desiderata for a good measurement: it should be reliable, eliminating or averaging random errors such that repeated measurements yield the same result; it should be valid, measuring the merit of research proposals rather than some other, perhaps more easily accessible, quantity. Chubin notes that survey respondents were concerned about both aspects of the measurement component of peer review: some worry that peer review measures prestige of applicants instead of the merit of their proposals, and others worry about a significant element of chance in peer review decisions.[23]

[20] This problem is related to the very fundamental problem in philosophy of science of knowing whether a research programme is progressing or degenerating, as noted by Lakatos (1970).

[21] Note that Chubin and Hackett's use of the term "rationality" is not identical to my own usage of the term, presented at the beginning of this section. In all places except this section, the term "rationality" refers to my own definition, unless stated otherwise.

[22] Detailed examples of the information flow from funding agency to scientists are given in NIH (2013b); NSF (2013b).

[23] Empirical evidence supporting the worry about a significant element of chance in peer review is presented in Chapter 6.

This list of desiderata emerges from a survey of practising scientists who have many years of experience with science funding by peer review. By comparing to other such field studies (Cole et al., 1977, 1981; Martino, 1992) we can safely conclude that Chubin's list is both comprehensive and relevant. However, it is clear that the different desiderata in the list do not stand on equal footing: some, e.g. effectiveness, are clearly more important than others. Furthermore, the different desiderata are not independent of each other. The following discussion charts the relations between them.

The first important set of relations is that surrounding the primary desideratum of *effectiveness*. Effectiveness itself is a somewhat elusive desideratum, as it takes its content from the general rationale for government support of science, which as we've seen in §1.3 could take different forms. Nonetheless, whatever content we fill it with, once a general purpose for science funding is decided, effectiveness is the pursuit of that purpose, and therefore takes primacy over other considerations. Four other desiderata are directly related to effectiveness: responsiveness, reliability, accountability, and rationality.

Both *responsiveness* and *reliability* are particular ways of promoting effectiveness; rather than tell us something above and beyond effectiveness, they sharpen our understanding of how effectiveness can be achieved. In the context of Chubin and Hackett's study, these desiderata also highlight what practising scientists consider to be the greatest challenges to effectiveness: the difficulty of striking a balance between conservatism and fashions, and the difficulty of reliably estimating the merit of proposed scientific projects.

Accountability and *rationality* are concerned, at least in part, with the appearance of effectiveness. One part of accountability is the appearance of effectiveness towards the public, who are providing the funds, and their elected representatives; the other part of accountability, arguably, is the ability of the public, via its representatives, to directly influence the operations of the allocation mechanism, which may be in tension with effectiveness. *Rationality*, as presented by Chubin, is the appearance of effectiveness, and transparency of operations, towards those participating in the system. However, rationality can also place a constraint on effectiveness, in case that an effective solution requires the deception or deliberate lack of knowledge among those participating in the allocation mechanism; in this thesis no such mechanism will be considered.

To summarise, from five distinct desiderata we are down to one, *effectiveness*, meaning the successful operation of the allocation mechanism in supporting valuable science. The four other desiderata sharpen our understanding of effectiveness, both in terms of its likely pitfalls, over- or under-conservatism and unreliable estimation of merit, and in terms of the need for the appearance of effectiveness alongside actual effectiveness, both towards the public and its representatives and towards the scientists participating in the allocation mechanism. At this stage effectiveness is still a rather vague desideratum; a more thorough analysis of the concept of the merit of a research project, from which follows a clearer notion of effective allocation, is presented in Chapter 3. For the moment, we can proceed with a general notion of effectiveness as the selection of valuable research projects, alongside the clarifications discussed above.

We are left with two other desiderata: *fairness* and *efficiency*. Both are in apparent tension with effectiveness, as they emerge from reasons other than those motivating effectiveness. *A priori*, there is no reason to think that a fair and/or efficient system will be effective; for example, distributing research funds equally to all members of society via a dedicated tax reduction would arguably be fair, but not effective. When considering efficiency, we may decide to give up a small

amount of effectiveness for a large saving in costs and time spent. However, a lack in fairness or efficiency may well result in reduced effectiveness, as scientists may refuse to participate in a system that is highly unfair or highly inefficient. Thus, we can think of efficiency and fairness as constraints on the design of an effective allocation mechanism: even if a particular mechanism seems highly effective, if it is grossly inefficient or blatantly unfair it will overall reduce the effectiveness of funding, and so any eventual allocation mechanism will need to fit between certain bounds of fairness and efficiency. Within these bounds, the optimal choice will be the most effective mechanism.

Now that we have a clearer notion of the desiderata a funding mechanism should meet, let us consider the extent to which the existing mechanism, peer review, meets these desiderata. In the sections below two positions are discussed: Polanyi's defence of peer review, and Gillies' critique of the practice.

1.5 A defence of peer review

The clearest defence to date of the practice of research funding allocation by peer review has been presented by Michael Polanyi. The outline of Polanyi's work is presented in three stages: §1.5.1 presents the historical context of Polanyi's argument;[24] §1.5.2 presents Polanyi's account of the concept of scientific merit; this account of scientific merit directly sets up the problem of science funding as one of balancing plausibility and originality, a problem which Polanyi argues is best addressed by a system of peer review, as presented in §1.5.3.

1.5.1 Historical background

In England in the 1930s a debate was ongoing between proponents of socially-relevant science, in the vein of Soviet science, on the one hand, and proponents of scientific freedom on the other hand. Naturally, this debate was embedded in the wider political and economic debates of the time, between Soviet-style communism on the one hand and free-market capitalism on the other. Two figures are emblematic of the narrower debate regarding science: Michael Polanyi and J. D. Bernal. Bernal, a marxist and a leading crystallographer who later played a role in the discovery of DNA (see Chapter 4), published *The Social Function of Science* (Bernal, 1939), in which he argued for a scientific establishment that is centrally directed towards meeting social needs. Bernal's book is composed of two parts: the first provides a survey of the historical relation between basic and applied science, and lists several sources of gross inefficiencies in the operations of the institution of science; in the second part, Bernal sketches a proposal for a more efficient institution of science, one modelled on Soviet science, where the aim of research is to alleviate immediate social concerns such as disease, hunger and unemployment. Admittedly, under Bernal's vision scientists take up this practically-oriented research of their own volition, once they are made aware of its social significance, so the vision does not involve a centrally-located cadre of bureaucrats assigning research tasks to scientific workers. However, a worry arises in case certain scientists do not wish to pursue the research Bernal has envisioned, and this worry is left unanswered.

[24] For a rich account of Polanyi's work and the conditions which led him to publicise the argument discussed in this section see Nye (2011).

In response to Bernal's book, Polanyi presented an argument in defence of "free science", which was later published in the first volume of the journal *Minerva* as *The Republic of Science* (Polanyi, 1962). In this paper, Polanyi argues for the appropriateness of two analogies: the first analogy is economic, between the "invisible hand" of the free market and the self-coordination of the scientific community to pursue at each stage the scientific problems of highest merit; the second analogy is political, between the independence of science and the independence of self-governing nation states. It is the former, economic, analogy which is relevant here, as it provides an argument regarding an effective method of choosing scientific projects, the primary desideratum of a funding mechanism. In contrast, the latter, political, analogy is of less interest to us here, since, as stated above and expanded in the next chapter, this thesis is concerned with the choice of projects once overall questions of the aim or value of research have been settled, regardless of whether the value of research is dominantly in its intrinsic appeal to the research community or in its applications.

1.5.2 Definition of scientific merit

To understand Polanyi's argument for effectiveness via peer review, we need to understand his conception of scientific merit. Polanyi asserts that three mutually-interacting factors contribute to scientific merit (Polanyi, 1962, pp. 57-8):

Plausibility A proposed scientific project needs to be plausible in light of previously held scientific knowledge. For example, research into extra-sensory perception or a novel perpetual motion device is unlikely to be attributed high merit, *ceteris paribus*.

Scientific Value The value of a project, which is only one contributing factor in its merit, is subdivided by Polanyi into three "coefficients":

1. Accuracy: the level of detail and probable error of the results. High accuracy, leading to greater value, means high level of detail with low probable error.
2. Systematic importance: the extent to which the results inform other areas, questions, or practices of scientific research. High systematic importance, and greater value, is assigned to projects that are relevant to many other projects.
3. Intrinsic interest of the subject matter: contemporary levels of lay interest in the subject matter of the project. Projects that address matters of high lay interest, e.g. by addressing emotionally charged topics or matters of economic importance, are assigned higher intrinsic interest and higher value.

According to Polanyi, the three coefficients can compensate for each other, so that a highly interesting but low accuracy project may be attributed sufficient value to warrant its research, and similarly for a highly accurate and systematically important project of less intrinsic interest.

Originality A proposed scientific project needs to be original and offer new knowledge or a novel outlook on existing experience. The greater the surprise caused by the research the higher its merit, *ceteris paribus*.

From the language of the description and examples given it is clear that Polanyi means this list to be *descriptive* of the evaluation assigned by practising scientists to proposed projects.

This is in keeping with his position that the choice of research projects should follow the internal judgement of the scientific community.

1.5.3 Argument for peer review

Polanyi's factors of scientific merit are set up to highlight an inherent tension: between plausibility and originality, between conservatism and innovation. A significant portion of Polanyi's paper is dedicated to this tension, as it explains the need for a strong scientific authority, while placing constraints on its form and mode of operation. It is also the central motivation for funding by peer review, as will be explained below. To make the tension more vivd, Polanyi outlines two extreme modes of organisation for research projects: one extremely-conservative, the other extremely-dispersed (the meaning of the latter will become clear shortly). The presentation of the extreme positions below is my own, adapted from Polanyi's text but not identical to his. In the text Polanyi moves back-and-forth between the different positions, and many of their characteristics are only implied in the context of a particular comparison to another setup. In the presentation below I have collected Polanyi's various points into three accounts (extremely-conservative, extremely-dispersed, and peer review) to bring out the differences between them. I then present Polanyi's own argument for why peer review is expected to outperform the two extremes, and in general function as the optimal mechanism for allocating science funding to projects of high scientific merit, given the definition of scientific merit presented above.

Extremely-conservative science funding

The extreme-conservative setup is presented by Polanyi as a criticism of narrowly-directed research. Implicitly, this is the position Polanyi attributes to Bernal and other proponents of Soviet science. I propose we envisage this setup in the following way. Assume that at a certain time t the totality of known scientific domains is $K = \{k_i\}$, and a central authority has decided to direct research towards a range of practical social needs $N = \{n_i\}$.[25] At that time, a certain mapping $M : K \to N$ is known from scientific domains to practical needs, that assigns to each domain of knowledge the areas of social need to which it is known to be relevant, e.g. from botany to food, shelter, and medicine (one can think of more or less detailed mappings). A strict emphasis on directing science towards meeting social needs would result in fostering only those fields of knowledge which map to non-empty sets of needs. The result of this investment will be increased theoretical advances in those fields, and increased yield of practical advances; it will not, however, result in new knowledge about the potential of the neglected fields, nor will it uncover new fields for research. In other words, the mapping M will not be updated. The only update to M will occur when the potential for practical payoff in a field is exhausted. Thus, over many iterations, all fields will map to empty sets of needs, and research will come to a halt.

This extreme-conservative setup can also be represented graphically, as in Fig.1.1.[26] The total area represents all projects that may be pursued. Red squares represent as-yet not pursued projects that would yield social benefit if pursued, white squares represent projects that have been pursued in the past, black squares represent as-yet not pursued projects with no associated

[25] In general, a central authority could direct research in any specific direction. In the context of Polanyi's paper the direction discussed is practical social needs, but this could be taken to be only one example of authoritarian direction of research.

[26] This pictorial representation is only a rough illustration of Polanyi's argument. A much more detailed representation, with some claim towards accuracy, is presented in Chapter 5.

social benefit. If at a certain time we narrow research around what are known to be socially-relevant areas (blue circles around clusters of red squares around white squares), and restrict all researchers to operate only in those areas, eventually all potentially-useful projects will be pursued within those areas (bottom panel), and research will come to a halt, even though there are still many unexplored projects of potential benefit (red squares).

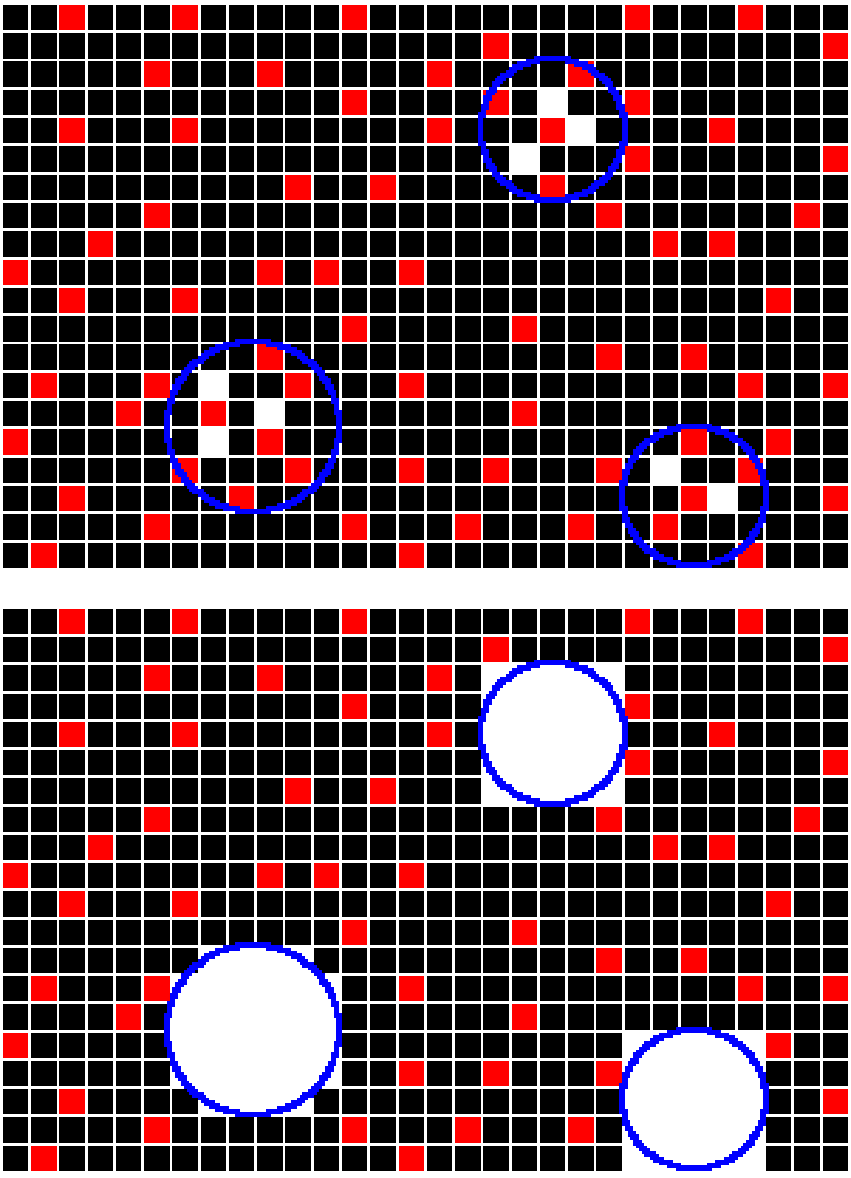


Figure 1.1: Illustration of Polanyi's argument regarding the progress of science under direction towards specific social needs. The grid represents all possible research projects, white squares are projects which have already been pursued, red squares are projects which, if pursued, will lead to social benefit, and black squares are projects which will not lead to social benefit if pursued. The top panel represents the state of research at a certain time, when direction towards specific needs is enacted by defining areas of special interest (blue circles) and restricting research to those areas. The bottom panel shows the situation after research has run its course: full exploration within these area and halt of scientific progress, while many potentially beneficial projects still exist.

Extremely-dispersed science funding

In contrast, the extremely-dispersed setup is provided by Polanyi to represent the state of affairs in the absence of scientific authority.[27] In this setup, each individual scientist pursues a project of her own choosing, regardless of the projects pursued by those that came before them or the choices of their peers. *Prima facie*, this would be the setup most likely to promote innovation, and if originality was the only measure of scientific merit, it would be the most effective. However, the potential for accuracy and systematic importance is reduced, as these often emerge from the interaction, and coordination, of many different avenues of research.

Furthermore, Polanyi argues that under this setup it will be nearly impossible to evaluate the plausibility of results. Science, Polanyi argues, is constantly besieged by "cranks, frauds, and bunglers" (p. 57). In order to tell apart the cranks from the valuable contributors, an authoritative body of knowledge is required. However, if all scientists are operating independently of the others, no such authority can emerge, or at least not one that is based on critical examination of empirical results.

The dispersed state will not last long, Polanyi argues, as the vacuum of scientific authority, combined with the need to identify the cranks, will give rise to an authoritative structure that is linked to political and economic power, to the detriment of the advance of science:

> In parts of the world where no sound and authoritative scientific opinion is established research stagnates for lack of stimulus, while unsound reputations grow up based on commonplace achievements or mere empty boasts. Politics and business play havoc with appointments and the granting of subsidies for research; journals are made unreadable by including much trash. (Polanyi, 1962, p. 61)

The extremely-dispersed setup can be represented graphically, as in Fig. 1.2. At any time an outsider is faced with a random selection of putative experts and contradicting claims to scientific results. In the absence of a scientific authority that can adjudicate the quality of results, the outsider is left with three options: accept all results, pick a set at random, or avoid all. Due to the high frequency of low-quality results (as there is no barrier to entry or policing of quality) the average quality is very low, and the rational course of action is to avoid relying on such a scientific industry altogether.[28]

[27] One can easily foresee Kuhn's notion of pre-paradigmatic science in this setup. The link between Polanyi's and Kuhn's work is explored by Nye (2011). This kind of "anything goes" setup is also reminiscent of the structure of science proposed by Feyerabend (1975).

[28] The topic of expert evaluation, both on an individual level and on a community level, is discussed by Kitcher (1993); Goldman (1999, 2001). This topic is somewhat tangential to this chapter, but is returned to at greater length in the next chapter.

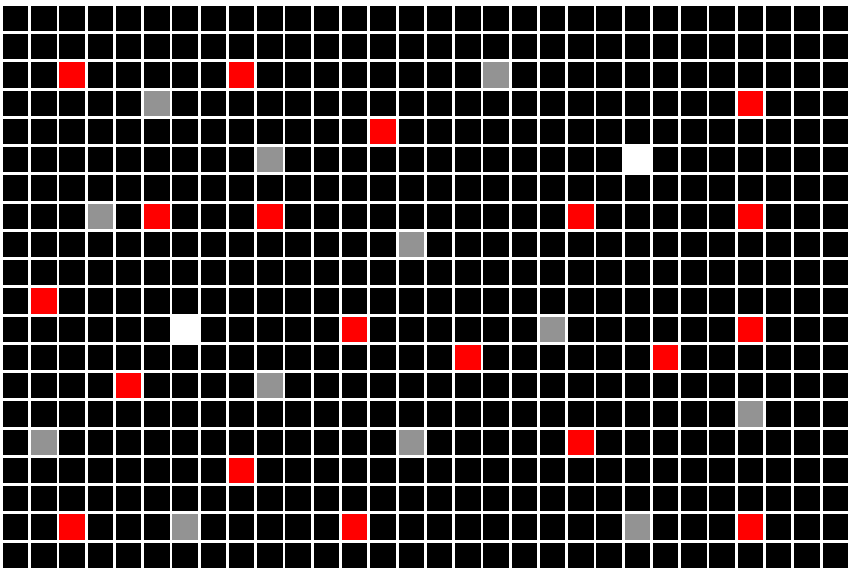


Figure 1.2: Illustration of Polanyi's argument regarding the state of science in the absence of coordination. The grid represents all possible research projects, colour of square presents status of project: (white) currently pursued and will lead to useful results, (grey) currently pursued but will not lead to useful results, (red) if pursued will lead to useful results, (black) will not lead to useful results if pursued. As there is no scientific authority to pass judgement, there is no way for an outsider to tell apart white and grey projects. Due to the high frequency of grey projects (cranks, frauds and bunglers, as well as simply unlucky scientists) the risk in accepting scientific results becomes too high, and the entire scientific industry becomes useless.

Peer review as optimal balance

Given the predicted results of the two extreme setups, it follows that the most effective setup of project choice lies somewhere in the middle: not too conservative, as it leads to increasingly diminishing originality, but not too dispersed, as it leads to very low plausibility. From his portrayal of the two extremes, it is clear that Polanyi places utmost importance on the role and nature of scientific authority in shaping the distribution of pursued projects. It comes as no surprise, then, that Polanyi's solution relies on the establishment of an optimal scientific authority. Since Polanyi sets out to defend the scientific institutions that were present at the time from attempted changes, his account blurs the normative/descriptive divide: he both describes the operations of the scientific institutions he sees around him, and argues for their superiority over other ways of organising science. Following the presentation of the two extremes above, and using Polanyi's list of the factors of scientific merit, we can reconstruct Polanyi's recipe for scientific authority in the following manner. Starting from the originality factor, we see that utmost originality is assigned to the extremely-dispersed setup, where individuals pursue topics of their own choosing. This setup however lacks an authoritative quality-check mechanism to ensure plausibility, accuracy and systemic importance. We add this quality-check mechanism in the form of a collective "scientific opinion" that passes judgement on proposed projects and published results. How could we instantiate such a mechanism? We would need to entrust these decisions to those individuals who are competent to make them. Who are these individuals? There is no person or small group with sufficient familiarity with all topics of scientific knowledge to make the necessary judgments; however, scientists who work on closely related projects can pass judgement on each other's work. Thus, we should maintain a mapping from fields of research

to scientists working in those fields. Any scientist would be free to choose any scientific project; however, her proposal and results must be checked and approved by those working in the field of her choice. This would guarantee the quality of research (plausibility and value in Polanyi's terminology), while maintaining sufficient originality. Thus, we arrive at a defence of free enquiry checked by peer review: it is the only mechanism that can simultaneously maximise originality, value, and plausibility.

In terms of the desiderata listed in the previous section, we can identify Polanyi's argument as defending peer review based on two aspects of *effectiveness*: *responsiveness* is guaranteed by the free choice of individual scientists, while *reliability* is guaranteed by using the only plausible evaluation mechanism: evaluation by those practising in the same field of inquiry. Of course, to simply state that this is the only plausible evaluation mechanism begs the question. Therefore, Polanyi provides the following argument:

> [E]ach scientist originally established himself as such by [...] [submitting] to a vast range of value-judgments exercised over all the domains of science, which [he] henceforth endorses, although he knows hardly anything about their subject-matter. [...] [T]he appreciation of scientific merit [...] is based on a tradition which succeeding generations accept and develop as their own scientific opinion. This conclusion gains important support from the fact that the methods of scientific inquiry cannot be explicitly formulated and hence can be transmitted only in the same way as an art, by the affiliation of apprentices to a master. (Polanyi, 1962, pp. 68-9)

There are two steps in Polanyi's argument. The first is descriptive: the kind of value-judgements required to evaluate scientific merit are spread throughout the community of scientists, and not outside it. These values are endorsed by individual scientists during and following their process of socialisation as scientists, but the entire range of these values is not explicitly available to any practising scientist. Rather, the entire set of value-judgements passes piecemeal from one generation of scientists to the next, and only the scientific community as a whole can be considered competent at passing judgment on questions of scientific merit. The realisation of this competence is in peer evaluation, where the most relevant values to the evaluation at hand reside. The second step of Polanyi's argument is that this state of affairs is *necessary*, because the set of values required to pass competent judgement cannot be explicitly formulated, and therefore cannot pass from one individual to another without a long process of apprenticeship. In other words, Polanyi argues that the knowledge required to make judgments of scientific merit is *tacit*, and will therefore only be available to those scientists whose training was similar to that of the scientist whose project is standing for evaluation, i.e. her peers.[29]

Polanyi's optimal solution can be represented pictorially, as in Fig. 1.3. Areas of accumulated knowledge and experience offer sufficient authority to allow evaluation of further work and growth. Work taking place outside of the umbrella of peer review cannot be evaluated, and is therefore ignored, even if it contains some potential merit.

Of course, those working in a certain field will need to be responsive to original contributions in their field, or the setup will devolve into the extreme-conservative setup. Polanyi argues that this kind of responsiveness is part of the socialisation of scientists and the norms they

[29] Polanyi has written extensively on the role of tacit knowledge in science (Polanyi, 1958, 1967). For a more recent account of tacit knowledge see Collins (2010).

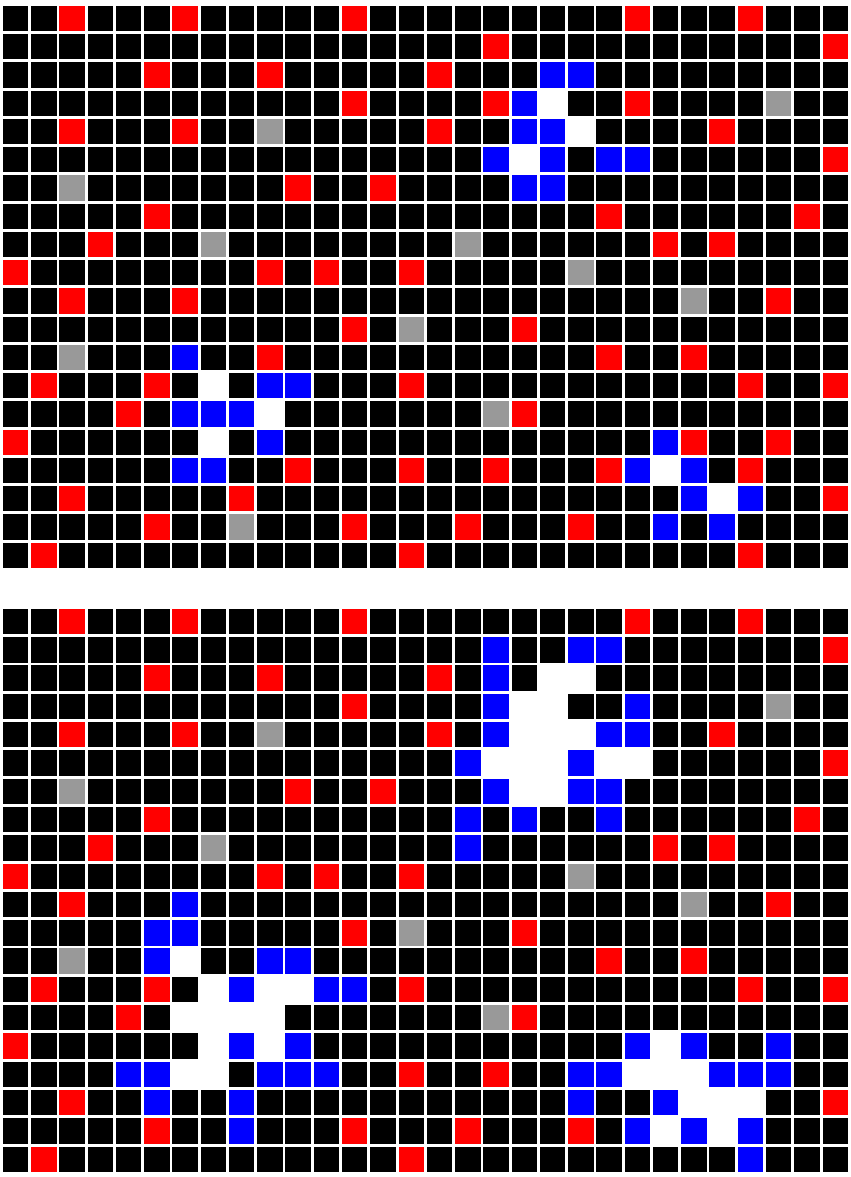


Figure 1.3: Illustration of Polanyi's proposed mechanism for scientific growth. The grid represents all possible research projects, white squares are projects which have already been pursued, red squares are projects which, if pursued, will lead to useful results, and black squares are projects which will not lead to social benefit if pursued. Blue squares represent currently pursued projects that fall within the scope of peer review, and therefore their results can be evaluated and ultimately incorporated into the scientific body of knowledge. Grey squares represent projects being pursued outside the scope of peer review, and are therefore ignored. The transition from top to bottom panel shows progress in time, depicting growth of knowledge around centres of scientific authority, while work outside of the umbrella of peer review contributes nothing.

exhibit. Furthermore, this kind of setup would prevent scientists from working in completely new, unoccupied, fields, that potentially hold much larger promise for novel contributions. This, Polanyi argues, is a cost that must be paid, invoking the threat of the extremely-dispersed setup:

> For scientific opinion may, of course, sometimes be mistaken, and as a result unorthodox work of high originality and merit may be discouraged or altogether suppressed for a time. But these risks have to be taken. Only the discipline imposed by an effective scientific opinion can prevent the adulteration of science by cranks and dabblers. (Polanyi, 1962, p. 61)

This last point seems to be the weakest point of Polanyi's argument, being little more than a slippery-slope fallacy. Would all admittance of unorthodox work necessarily lead to the chaotic picture of "anything goes"? Gillies argues that this is not the case, and offers a critique of peer review based on this point, as discussed below.

1.6 A critique of peer review

In a series of papers and a book, historian and philosopher of science Donald Gillies has presented a critique of grant peer review. Gillies uses historical examples to highlight important failures of peer review, and then provides explanations for these failures using philosophical analysis, mostly drawing on Kuhn (1996). The juxtaposition of Polanyi's arguments with Gillies' arguments provides extensive philosophical expansion on issues highlighted by Chubin and Hackett, mostly on tensions that emerge in efforts to balance responsiveness and reliability, and responsiveness and accountability. The different ways in which these tensions can be resolved have direct influence on the overall effectiveness of the allocation mechanism.

The presentation of Gillies' work on peer review is divided into three parts. The first part presents an overview of Gillies' criticism of peer review on the grounds that it suppresses innovative research and heterogeneity of research programmes (Gillies, 2007, 2008, 2012, 2014). The second part presents Gillies' proposals for alternative funding mechanisms, one that supports all those interested in conducting non-laboratory academic research (Gillies, 2008, 2009) and one that distributes funds by lottery in the case of laboratory research (Gillies, 2014). The third and last part presents my own evaluation of Gillies' arguments and proposals, and outlines its relevance to the remainder of this thesis.

1.6.1 Peer review suppresses innovation and heterogeneity

In 2008 the UK government retired the Research Assessment Exercise (RAE), a peer-review based funding mechanism for academic research, in favour of a new mechanism called the Research Excellence Framework.[30] Designed to coincide with the heightened public debate surrounding the funding of higher education, Gillies published a series of papers (Gillies, 2007, 2009) and a book (Gillies, 2008) in which he criticised the RAE, and peer-review based funding schemes in general.[31] The main line of Gillies' criticism is that peer-review is inherently biased towards poor

[30] To be precise, only one funding stream was determined by the RAE, and that was the funding provided by the Higher Education Funding Council for England (HEFCE). Outside of HEFCE, government research funds in the UK are also provided by the different Research Councils, which operate in a manner more similar to the NIH and NSF, summarised in §1.2.

[31] While Gillies mostly targets a specific implementation of peer review, his argument targets fundamental aspects that are common to most peer review mechanisms (See §1.2). An important difference between the RAE

evaluation of highly innovative, and meritorious, research, and therefore it is not an effective way of allocating research funds.

Gillies' argument is presented below in four steps, roughly following the order in which they are presented in Gillies (2007, 2008). The first step presents historical evidence that highly innovative researchers have indeed suffered from poor peer assessment of their early work, and were only recognised for their remarkable contributions later on, after managing to support and publicise their research via means other than peer-review based evaluation. To these cases I add the cases presented in Gillies (2012, 2014) that demonstrate peer review's suppression of heterogeneous approaches.

The second step presents a possible explanation of the phenomenon of poor evaluation of highly innovative research by drawing on Kuhn's notion of a scientific paradigm. The third step argues for the applicability of the Kuhnian framework to the historical cases considered in the first step. The fourth step generalises from the explanations to a general criticism of peer review based on its suppression of innovation and heterogeneity, and the overall reduction in beneficial scientific outputs that follows from this suppression. Note that this general argument is in direct opposition to Polanyi's argument, where suppression of innovation is necessary for supporting and maintaining the generation of beneficial scientific outputs.

Historical cases

Gillies' first test case is Frege.[32] In 1879, Frege published *Begriffsschrift*, a book presenting a novel approach in mathematical logic involving an axiomatic-deductive approach to the calculus of propositions and predicates. Frege today is considered a key figure in the foundation of modern mathematical logic, and much of his original contribution is still maintained in contemporary textbooks of mathematical logic. Gillies presents a list of highly favourable reviews of Frege's work from eminent logicians of the 1950s and 1960s, calling Frege *The* father of modern logic and comparing the importance of his work to that of Aristotle's. Gillies juxtaposes these reviews with the reviews of Frege's work by his contemporaries, collected shortly after the publication of *Begriffsschrift*. These reviews label Frege's work as unoriginal, as detracting from the accumulation of knowledge in logic, and criticise it for presenting unnecessary new notation and language that achieves little. Clearly, if Frege's work would have required approval by his peers for its support, we would have been denied this remarkable achievement, which later had great applied importance in the field of computation.

The second test case is Semmelweis. Semmelweis, while working as a medical doctor in a maternity ward in Vienna from 1844 to 1849, discovered by a series of hypotheses and tests that "cadaverous material" from autopsies, transferred to ward patients despite hand washing in soapy water, was the cause of increased mortality rates. Semmelweis' discovery is vindicated by modern understanding of the situation, attributable to certain kinds of bacterial infection. Modern practice also vindicates Semmelweis' recommendation, that doctors wash their hands in a stronger disinfectant after autopsies. However, Semmelweis' contemporaries rejected his

and the funding mechanisms discussed is that the RAE exercises *ex post* evaluation of past projects to determine funding of future projects, rather than the more common *ex ante* evaluation of proposed projects based on a research proposal. However, in the context of Gillies' argument this makes little difference, as Gillies centres his argument on the unreliability of the evaluation process; it naturally follows that if an *ex post* evaluation is unreliable, an *ex ante* evaluation of the same quantity would be as unreliable or more so, since less information is available to guide the evaluation.

[32]The first three test cases are presented in Gillies (2007, 2008).

explanation and recommendations, and antiseptics in hospitals were not introduced into general use until 20 years later. Gillies claims that if Semmelweis' investigations would have depended on approval by his peers, his discoveries would not have been made, and the introduction of antiseptics would have been delayed even further.

Gillies' third historical case is Copernicus, which further exemplifies the general theme of revolutionaries being rejected by their contemporaries only to be later vindicated and hailed as great innovators. In these three cases, if we imagine a counterfactual historical RAE, the results are grim, as we would have lost highly valuable research due to poor peer evaluation.

To these three historical cases of poor initial peer evaluation Gillies (2012, 2014) adds two more recent cases that exemplify general rejection of innovation or suppression of heterogeneity by peer review.

Gillies (2012) argues that in the field of economics, the RAE has contributed to a high level of homogeneity via the support of neo-classical economics as the orthodoxy of the field and the suppression of "heterodox economics" schools such as Marxist economics or post-Keynesian economics. Evidence for the effect of the RAE on heterogeneity can be seen in the data collected by Lee (2007). Lee compared data from the departmental 2001 RAE scores to data on the venues of publication for economists in those departments. In particular, Lee was interested in the percentage of publications in the "diamond list", a list of 27 journals which publish almost exclusively papers in mainstream neo-classical economics, and the percentage of publications labelled by the RAE evaluators as "Heterodox, History of Economics Thought, and Methodology" (H-HET-M). The data show very clear correlations: the higher the percentage of "diamond list" publications, the higher the RAE score, and the higher the number of H-HET-M publications, the lower the RAE score. Gillies contrasts these correlations, that suggest neo-classical economics is somehow "better", with evidence that in terms of making predictions regarding the 2008 financial crash, neo-classical economics did no better, and in fact worse, than heterodox economics. Gillies gives examples of leading neo-classical economists, such as Robert E. Lucas and Richard Fortes, who directly addressed relevant economic issues and risks such as the solvency of Icelandic banks, and yet failed to predict the 2008 financial crash. Gillies contrasts them with examples of heterodox economists working on the problem, such as Steve Keen, who did successfully predict the crash, though unfortunately their warnings were not headed, partly due to the status of neo-classical economics as orthodoxy.

The last test case is zur Hausen, who in 2008 received the Nobel Prize for discovering a relation between the papilloma virus and cervical cancer. In the 1970s zur Hausen's approach was in the minority, as prior discoveries linking other forms of cancer to herpes virus led the majority of researchers to believe cervical cancer will also be related to herpes virus. In 1972 a major conference dealing with the topic of cervical cancer was held under the title "Herpesvirus and Cervical Cancer". In this conference zur Hausen presented a criticism of the herpes-focused approach, but was met with "stony silence", and his contribution was not mentioned in the conference summary. Zur Hausen's research was supported by his university in Germany, and he therefore did not need to apply for peer review. Gillies argues that had he needed to apply for prospective peer review to fund his research his application would have been given a very low score, and the research, and the resulting discovery, would not have taken place. Zur Hausen's discovery led to a development of a vaccine which prevented, every year since vaccines became widespread, many thousands of cases of cervical cancer, and generated multi-million pound

revenues for GlaxoSmithKline, benefits which would not have been generated under the peer review system.

Explaining the historical cases

Why is it that highly innovative research, identified as such several decades after the initial contribution, initially receives poor peer evaluation? According to Gillies, this can be explained by using the Kuhnian notion of a paradigm.[33] A paradigm, very roughly, is an accepted way of defining the legitimate questions in a field of research, and of evaluating proposed solutions to these problems. A paradigm enables a community of investigators to develop and accumulate a large amount of highly specialised knowledge in a specific domain of inquiry, though it often leaves potentially interesting questions outside the "legitimate" domain of inquiry as mere "anomalies". According to Kuhn, science makes progress through alternating phases of "Normal Science", during which a dominant paradigm is elaborated and expanded, and "Revolutionary Science", during which an established paradigm is replaced by a new one following a crisis of the old paradigm brought about by the accumulation of too many anomalies.

Most of the time, science is found to be in the "Normal" mode, and the majority of researchers follow the dominant paradigm. A significant amount of time has already been spent investigating the domain under the guidance of the paradigm, and therefore highly innovative research within the confines of the paradigm is unlikely. This means any highly innovative research is likely to occur outside the focus of the paradigm, with the innovator being a "revolutionary". Since the paradigm defines the range of legitimate questions and admissible answers, and since the peers of the revolutionary are likely to be adherents of the paradigm, their assessment of the revolutionary's work will likely be poor, leading to the observed historical phenomena detailed above.

Applicability of the explanation

Gillies' explanation of the poor assessment of revolutionary research caused by paradigm-based science would account for the first three historical examples, *if* it can be shown that Frege, Semmelweis and Copernicus were indeed engaged in revolutionary science, ushering in a new paradigm. For the Copernican case this is fairly straightforward, as Kuhn himself uses the Copernican Revolution as an example of a paradigm shift (Kuhn, 1996). The applicability of Kuhn's framework to Frege and Semmelweis has been argued for by Gillies in earlier work: Gillies uses the case of Frege to argue for the applicability of the notion of revolution to mathematics (Gillies, 1992) and the case of Semmelweis to argue for the applicability of a (slightly altered) notion of paradigm to medicine (Gillies, 2005). These works are rich in historical detail and philosophical analysis, and it is reasonable to accept that the Kuhnian framework can indeed be successfully applied to the first three test cases presented by Gillies.

For the last two cases Gillies does not provide a Khunian revolutionary-science account. For the case of economics, Gillies follows Kuhn's position that the social sciences contain multiple, competing paradigms, and that followers of one paradigm tend to have a lower evaluation of

[33]The concept of a scientific paradigm is presented in Kuhn (1996). Following the publication of the first edition of his book in 1962 a conference was held in LSE to discuss Kuhn's work in relation to the work of Popper; the papers from the conference were published in Lakatos and Musgrave (1970). One paper of particular note is by Masterman (1970), analysing the different ways the notion of "paradigm" is used in Kuhn's book. For a recent account of Kuhn's philosophy see Bird (2000).

those following a different paradigm. It is clear from the data presented by Lee (2007) that school affiliation certainly plays some role in peer review. In the case of zur Hausen's research, Gillies provides an account based on the notion of a research programme presented by Lakatos (1970), with the expectation that followers of different research programmes would make lower evaluations of those following a different programme.

General criticism of peer review

Once established as a successful explanation of the historical cases, Gillies generalises from the explanation to a criticism of peer review. The effectiveness of peer review, Gillies argues, hinges on its *reliability* as a measurement of good research. *Qua* measurement, the reliability of peer review should be estimated via four parameters: the cost and the likelihood of Type 1 and Type 2 errors. Type 1 errors occur when good research projects are mistakenly classified as bad research and do not get support; Type 2 errors occur when bad research projects are mistakenly classified as good research projects and get support. Generally, in the design of a measurement, there is a tradeoff between minimising Type 1 errors and minimising Type 2 errors.

Using the Kuhnian framework, Gillies offers an estimation of the likelihood and magnitude of both types of errors. Type 2 errors are likely to be rare under a successful implementation of peer review, because reviewers can rely on the paradigm and its past successes to reliably identify projects that, in light of the paradigm, would not lead to significant results. However, following the line of argument in the explanation of the historical cases, very costly Type 1 errors are likely to occur under peer review, because highly innovative, and therefore highly valuable, research is likely to be unreliably classified because the paradigm fails to guide (and in fact misguides) in its evaluation. The risk of committing this kind of Type 1 error, which would throw away some of the most successful of scientific discoveries, is intolerable, and makes peer review a non-viable funding mechanism for research.

In the case of economics, Gillies (2012) argues for a worrying global effect of peer review. Over time, peer review will tend to make a research community more homogeneous, suppressing minority schools which have a different set of tools and assumptions. Based on the evidence presented, it is clear that at least in economics, a significant portion of the merit score assigned by researcher A to researcher B depends on the schools A and B belong to. This high level of subjective variability is worrying for a measure that is used to compare candidate projects across the discipline. Furthermore, the variability is not random, but a systemic bias: cross-school evaluations are systemically biased negatively, and apparently same-school evaluations are systemically biased positively. This bias suggests a significant worry about the ability of peer review to be *responsive* to good ideas coming from minority schools.

Gillies (2014) presents a slightly different argument from the ones presented above. The argument is developed from a critique of peer review written by Sir James Black, Nobel Laureate and discoverer of two blockbuster drugs.[34] Black argues that the slowdown in the rate of discovery of new drugs in the beginning of the 21st century was not due to "the culling of low-hanging fruit", but rather due to a failure in the system of peer review, a system that, according to Black, destroys scientific creativity:

> The peer reviewers prefer the security of the well-advanced application to the specu-

[34]The paper, titled *The Business of Science in the Pharmaceutical Industry* was written in 2009, but was never published due to Sir James Black's death in 2010.

> lative, they prefer the group to the individual, they prefer the fashionable to the new. ... It is the prospective nature of the review, carried out anonymously, that is the problem. (Gillies, 2014, p. 4, quoted from p. 5 of Black's unpublished manuscript)

The anti-creative prejudice of peer review, Black and Gillies argue, is believed throughout the research community, leading researchers with creative ideas to keep silent about them and instead seek funding for "safer" projects. This effect is labelled by Black as "undesirable feedback".

Gillies argues that the failure of peer review in supporting creative research stems from "researcher narcissism" (Gillies, 2014, p. 8). Under this condition, individual researchers believe their chosen approach to the topic is the best possible approach. The first reason that "researcher narcissism" emerges is psychological. Each individual researcher spends a lot of time choosing their approach, and in the process of choosing they come up with reasons for justifying their eventual decision. These reasons are likely to be shared by other researchers surrounding the individual, as individuals choose a research community that will support their approach. This echoing of reasons further entrenches the individuals' belief in the justification of the superiority of their chosen approach. The second reason for the emergence of "researcher narcissism" is personal benefit. Individual researchers will enjoy more successful careers if others, including funding bodies, awarding bodies, and students, choose to endorse or follow the same approach as the individual researcher. Conversely, if the field rejects the approach of the researcher, negative career consequences are likely to follow. This incentive to have one's own approach succeed will often lead individuals to believe that it will indeed succeed, or at least behave as if this was their belief.

While it is clear that in a heterogeneous research community not all researchers can be justified in believing their chosen approach is the best one, Gillies goes further to argue that *virtually all* researchers are mistaken in their "narcissist" beliefs. This is "because no one really knows in advance what research projects are going to succeed. This is because research is, by definition, the exploration of things, which are as yet unknown" (Gillies, 2014, p. 7). This point is worth noting as it will return, in much greater length, throughout this thesis, as discussed in the evaluation of Gillies' argument below.

1.6.2 Proposals for alternative funding mechanisms

Following his criticism of peer review, Gillies (2008, 2009) advocated an alternative mechanism of "funding all scholars". In these works he restricts the proposal to non-laboratory research, and therefore this proposal has only limited, but interesting, overlap with the work presented in this thesis. In a recent paper Gillies (2014) proposed that research funds should be distributed, in the laboratory-science context, using random selection. A similar proposal is made in this thesis, though there are some important differences between this thesis' arguments for the benefits of random selection and Gillies' argument, and also between this thesis' recommendation for an alternative funding mechanism and Gillies' proposed random selection.

Fund all in non-laboratory research

Under Gillies' "fund all scholars" mechanism, all scholars actively pursuing research are given a stipend, and they are allowed to pursue any research topic without having to justify their choice to a peer-review body. Such a solution will clearly have very low costs for administration and

participation (no need to write and review grant proposals), but will it be effective in producing high-quality research? Gillies provides the following further specifications and arguments in its defence:

Retain peer review for appointments and promotions Gillies argues that in the context of appointments and promotions (once every several years) there is sufficient accumulated output to rely on peer review, and that in these contexts the ills of peer-review are less relevant (Gillies, 2008, pp. 95-99); while peer review may stop revolutionary scholars from gaining promotions, they would still be able to carry out their innovative research under a lower pay grade. This places one stage of quality assessment outside the funding mechanism, by relying on faculty admission procedures to serve as a filter of individual research capability. Such filtering would serve to limit the occurrence of Type 2 errors that result from general incompetence.

Easy transfer to teaching and administration Gillies considers the occupational landscape for scholars as including three major loci: research, teaching, and administration. Unlike research, teaching and administration have relatively-straightforward methods for evaluating contributions. If contributions were remunerated accordingly, and offered a reasonable income, many scholars, Gillies argues, will voluntarily transfer from research into teaching or administration, assuming such transfer was organisationally supported. Such transfers would ease the burden of supporting research, by allowing self-selection to limit the number of the less-qualified, or the no-longer-qualified, who are "stuck" in research.

Evaluation of industry The evaluation of industry, e.g. time dedicated to research or quantity of notes produced, is not riddled by the uncertainties and difficulties of the evaluation of merit, and would therefore reduce wasteful Type 2 errors without an intolerable decrease in efficiency. Legal requirements against external sources of funding or remuneration would further make sure academic research is not a haven for researchers not genuinely interested in research.

Fund laboratory research using random selection

Given Sir James Black's criticism of peer review, and Gillies' own example of zur Hausen's discovery of the role of the papilloma virus in cervical cancer, Gillies argues that peer review is not a good system for allocating funds to laboratory research. However, the high costs of laboratory research prevent the use of the "fund all scholars" mechanism in this setting. As an alternative, Gillies proposes selecting which researchers to fund based on random selection. Proposals will still be generated by individual researchers, as they are under peer review. Upon receipt of proposals by a funding body employing random choice, each proposal will first be checked for two things:

1. The funding body will check that the proposed research falls within the field of the grant.

2. The funding body will check that the individual applying for the grant is competent in the field of research suggested. This could be achieved by checking for qualifications such as a research PhD, experience as a research assistant in other projects, or by the applicant's track record of research and publication.

This checking, Gillies argues, would be relatively straightforward and much cheaper than peer review, while still addressing the need to filter out "cranks", thus implicitly addressing Polanyi's worry about the overwhelming number of low quality research projects that will emerge in an extremely-dispersed setup. Once proposals have been checked and filtered so only "serious" proposals are considered, a random selection (e.g. using a die or a lottery machine) is performed from amongst these "serious" proposals to select those which will receive funding.

While the immediate response is to think that random selection will be less effective than peer review in supporting good research, Gillies argues that the converse is true, due to the widespread systemic bias resulting from "researcher narcissism" as described above. Furthermore, it is envisaged that once peer review is removed applicants will have no reason to opt for "safe" proposals, and the level of innovation in proposals will increase, thus leading to more creative and beneficial research. Gillies proposes that random choice may first be introduced in a restricted set of institutions, and research can then be carried out to monitor its performance.

1.6.3 Evaluation of the critique

Gillies' arguments provide serious cause for concern regarding the operation of peer review. Based on the evidence presented, we can reject Polanyi's notion of a single "scientific opinion" shared by all practising scientists, at least in some cases. There is not one uniform set of criteria with which to evaluate the merit of projects and results, but schools or paradigms that offer sets of different criteria, and these sets of criteria may lead to contradictory evaluations. Using multiple examples, Gillies shows that the dominant set of criteria at a given time can turn out to be quite wrong in its evaluation of merit, and that a set of criteria endorsed by highly innovative individuals or minority schools of research can sometimes provide a more accurate evaluation of merit, as seen by later vindication. Gillies shows that the notion that peer review generates an accurate and objective measure of merit is untenable. Furthermore, peer review gives room to specific biases relating to contingent factors, such as dogmatic adherence to a paradigm or research-school affiliation, or to a researcher's own past and present decisions, to play a predictable negative role in the advancement of science. The counterfactual rejection of highly innovative research, and the evidence of the detrimental suppression of heterogeneity, show these are not fringe effects, but serious concerns for the efficient generation of scientific progress.

However, two general worries about Gillies' arguments should be mentioned. The first worry is about the universality, or lack thereof, of the negative effects described. As described by Gillies (2008), Wittgenstein's early work was positively evaluated by eminent peers despite its highly innovative nature, and there was at least enough support for heterodox economics for it to produce detailed warning about the coming financial crisis. Furthermore, it would be impossible to argue that ever since the rise to dominance of peer review as the funding mechanism for basic science no highly innovative research was supported. Since Gillies' argument of poor evaluation does not apply to all cases of peer review, we need to evaluate the degree to which his proposed mechanisms perform better than peer review overall: are the negative effects of peer review worse than the effectiveness benefits of selecting the most meritorious from each batch of proposals? Gillies' examples function well in highlighting specific failures of peer review, but it could be argued that they all lie on an extreme end of a spectrum, and that most of the time peer review is actually pretty good. The examples themselves do not offer a way to assess the matter of degree. Gillies' reliance on the Kuhnian framework suggests this might be a potential route to

a universal assessment of the issue, but Kuhn's notion of a paradigm is notoriously fuzzy, and therefore the road to generalising the concern relating to paradigm-dependant bias is not at all clear. A different general argument is presented in Gillies' account of "researcher narcissism", arguing that the merits of future research are inherently unknowable. This thesis develops this sceptical argument as the main criticism of peer review, which is only stated, rather than argued for, in Gillies' recent paper.

A second worry, relating to the first, is that Gillies only focuses on one aspect of scientific merit, and that is the bias introduced by paradigms. The extreme case of this is apparent in his model of peer assessment (Gillies, 2012), where economist A gives economist B a score of 1 if they belong to the same school, and a score of 0 if they do not. Gillies provides evidence that such bias does exist to varying extents, and sometimes plays a significant role, but that cannot be the whole story of merit evaluation. However, Gillies provides very little in accounting for the other factors of merit and its evaluation, such as the plausibility of the proposed project, the significance of its results (if successful), etc. It is these other factors that play an important role in Polanyi's defence of peer review, and their absence from Gillies' work raises concerns about lack of completeness in his treatment of the topic. A fuller account of merit, that considers both Gillies' familiarity-related bias and Polanyi's component-analysis of merit, is required, and such an account is developed in later chapters of this thesis.

An important lesson from Gillies' arguments and proposed solutions is that a funding mechanism can be inefficient in more than one way. There are differences between researchers who selfishly draw an academic salary while doing no work, researchers who slowly produce low quality work because they would rather be teaching but are "stuck" in research due to perverse incentives, researchers who falsely believe they are the next revolutionaries of their field, and researchers who are simply unlucky in their experimental setup, even though in all cases the result is the same: they received money that could have been used to support higher-quality scientific outputs. The differences lie not just in our moral judgement of these different causes of inefficiency, but also in the kinds of mechanisms that can be used to address them. Multiple mechanisms for promoting high-utilisation of research funds can be used across a wide range of research management institutions, some of which may operate outside the narrow operation of the institution responsible for directly allocating research funds. For example, encouraged self-selection or monitoring of industry produced can operate after funds have been allocated, as shown in Gillies' proposed mechanisms. This observation will be relevant when I propose my own alternative funding mechanism in Chapter 6.

The mechanism proposed by this thesis for allocating funds to research incorporates a significant element of random selection (details are given in Chapter 6). While others have proposed the consideration of introducing a formal random element into grant selection in the past (Greenberg, 1998; Ioannidis, 2011; Graves et al., 2011), the proposal presented by Gillies (2014) is closest to the one recommended by this thesis. This is not surprising, as both Gillies' proposal and my own emerge as a response to a worry that the relative merits of different research projects may be impossible to evaluate *ex ante*. However, Gillies and I develop different arguments with different focuses. As discussed above, Gillies offers two different arguments for bias in merit evaluation: the first argument argues for bias as a result of paradigm-dependant evaluation of merit, and includes epistemological, methodological and psychological aspects, but only applies when the reviewers and the scientists being reviewed pursue different paradigms or belong to

different schools; the second argument applies more globally, and argues for a psychological bias, "researcher narcissism", which hides from reviewers the true unpredictable nature of research, though this unpredictability is not argued for. My argument complements Gillies' second argument, providing epistemological justification for the unpredictability of research globally, i.e. even in cases when the reviewers and the scientists being reviewed pursue the same overall paradigm. However, I also investigate the sources and boundaries of unpredictability. This investigation leads me to propose, in contrast to Gillies, a more balanced mechanism that combines peer evaluation with random selection, and also enables me to suggest various circumstances in which random selection would not be expected to yield good results. In this way, this thesis combines elements from both Polanyi's and Gillies' proposals, and does so by developing an argument which has a broader focus and rests on less contestable foundations than Gillies' and Polanyi's arguments.

Conclusion

Grant peer review is the dominant allocation mechanism for public funding of basic research. It relies on two core commitments: that the public should support basic research, and that choice of research projects should be based on the project's evaluation by the peers of the project's proposing investigator. The justification for the first commitment, that first initiated substantial government support for basic research, comes from the expected social benefit of investment in research, in terms of health, economy, quality of life, and defence. The justification for peer evaluation is that scientific peers are in the best possible position to make an expert evaluation regarding a project's merits.

Since its initiation, peer review has become widespread and institutionalised in government bureaucracies. Operation over decades has brought a better understanding of the different interest groups involved in peer review, and their different expectations of peer review. In general, we can say that a funding mechanism should be effective at supporting high quality research, be seen to do so from within and without, and operate between certain bounds of efficiency and fairness.

The most worrying criticism of peer review is that it is overly conservative and risk-averse, that it stifles high-value innovative research in an unnecessary and costly crusade against low quality proposals. While Polanyi argues this conservatism is necessary in order to maintain the authority of science, Gillies argues this conservatism runs the risk of significantly slowing scientific progress, and proposes a radically different mechanism of funding all scholars in non-laboratory science, and funding laboratory science using random selection. Both Polanyi and Gillies discuss some aspects of scientific merit, while ignoring others. The next chapter argues that a model-based approach can provide a more comprehensive treatment of scientific merit and its evaluation in the context of science funding.

Chapter 2

Existing Models of Project Choice

Science is more like
climbing mountains in a fog
than finding lost dogs.

Introduction

The previous chapter introduced grant peer review, the dominant mode of contemporary allocation of public funds to basic research projects. To further develop the critique of peer review presented in the previous chapter, over the following chapters a strategy is adopted by which novel theory is developed alongside the presentation of historical examples, with new models offering mediation between theory and historical cases. These models are then used to inform the design of an alternative funding mechanism.

As mentioned in the introduction, the strategy pursued in this thesis is guided by the following question:

Are the processes used by public science funding bodies to make funding decisions rational, and can they be made more rational?

As indicated in the previous chapter, the general topic of science funding is too broad to be effectively considered as a whole. In §2.1 the question highlighted above is split into four questions, each at a different level of granularity, with a corresponding field of literature and approaches. This thesis will focus on the most fine-grained level of analysis, i.e. at the choice of individual research projects from a set of alternative proposals, and so treatment of existing literature will be focused accordingly on research-project choice. The existing literature on research-project choice has made significant use of *models*. §2.2 considers and justifies a model-centred approach for the project presented in this thesis, given difficulties in collecting novel empirical data in the domain and given the particular usefulness of models in institutional design and institutional critique.

§2.3 and §2.4 present a brief history of influential models of project choice. While they do not address directly the topic of science funding, these models provide a useful context and relevant conceptual tools for developing models that can address the topic of this thesis. §2.3 outlines the models of project choice presented in the works of Peirce (1879/1967), Kitcher (1990, 1993), and Strevens (2003). The section highlights some of the common assumptions made in these models, and presents the criticism of these assumptions given in Kitcher (2001) and Muldoon and

Weisberg (2011), as well as my own criticism. My criticism goes beyond Kitcher's criticism and Muldoon and Weisberg's criticism in pointing out that merit assignments, which in the models are presented as fixed and single-valued, are likely to both vary within a population and change over time.

§2.4 discusses the model of project choice presented by Weisberg and Muldoon (2009), which was presented by the authors as a direct response to the criticism of the "first-generation" models of Kitcher and Strevens. While providing important improvements over the first generation models, Weisberg and Muldoon's model does not address the above mentioned worries about differing merit assignments and time-variable merit. Furthermore, I argue that the scope of their model is too limited for the purpose of this thesis, because it only looks at processes *within* the scientific community. However, in order to represent science funding decisions, a model needs to include the wider scope of science-in-society.

The discussion of existing models of project choice, and their deficiencies, sets the scene for the next chapter, where alternative models of project choice are presented.

2.1 Survey of possible approaches

There are many questions that can be asked, and have been asked, about the choice of research projects – examples are given below. These questions can be separated, to an extent, into different loci of analysis, as they differ in their understanding of the topic and its pertinent questions, and in their approaches to answering those questions. To make some headway in surveying this field and locating the specific set of questions addressed by this thesis, we can distinguish four levels of decision-making, which create four, partially overlapping, loci of pertinent questions. From the most coarse-grained to the most fine-grained, these are:

1. How should a society, or its representatives or rulers, choose general areas of human interest, and prioritise their investigation? For example, should the society prioritise defence research or health research?

2. How should a society, or its appointed administrators and legislators, design the mix of institutions and practices that will distribute available funds in a certain area of interest? For example, should funding in the field of health be controlled by government or dominated by industry? Does the field of defence require a central coordinating institution, or will competition between institutions lead to better results?

3a. How should specific funding bodies, established to address a particular field or range of fields, choose the individuals and groups to be funded, and which methods should they use to select those individuals or groups? For example, should funding bodies seek experience or creativity? Should abilities be evaluated using a written proposal or an interview?

3b. How should funding bodies, or funded scientists, choose the research projects an individual or group will work on? For example, how should one estimate what is a reasonable goal in the field given the available funding? What principles should guide the choice of methods to pursue the chosen goal? What is the best way to deal with uncertainties?

The last two items on the list are not always at different levels of granularity, and are often considered together, for example when a funding body decides whether a *particular scientist*,

who submitted a *specific research proposal*, should be funded. Nonetheless, the questions at this most fine-grained level can be separated into two different loci, as sketched above, one focusing on scientists and the other focusing on research projects. Each of these loci can be addressed, to some extent, separately.

For each of the four levels there is a field of human sciences most appropriate for dealing with it:

1. The most general level deals with setting societal goals, which require (at least in a democratic society) the aggregation of many individual preferences into a plan for collective action; this is a political problem, to be tackled using political science and political philosophy. Kitcher's ideal of well-ordered science (Kitcher, 2001, 2011) is exemplary of such work.

2. At the level of mix of institutions, under the assumption that the political question is settled, the question bears on the most efficient organisation of humans to carry out the collective plan, on the most efficient flow of funds through the system, and on the social processes that may encourage or hinder institutions from carrying out their task. This is best tackled using sociological and economic tools. The sociological and economic articles in Geuna et al. (2003) serve as a good recent example of such work.

3a. At the level of individuals and groups, the difficulty is both to understand and measure the abilities of researchers, and to understand the ways to incentivise researchers to carry out the collective goal in the most efficient manner. These difficulties are best treated using methods from psychology and management theory. Some survey-based studies are discussed in Chubin and Hackett (1990); Martino (1992).

3b. At the level of the individual research project, the difficulty is in locating the project in the field of knowledge, predicting its outcomes, and choosing the methods most appropriate for use. At a practical level, the challenge is addressed by relying on scientific knowledge and insight from the history of science. At a general, meta-analysis level (the one adopted in this thesis), these are difficulties about what can be known and how to gain knowledge, to be tackled using epistemology and theories of scientific methodology. Some important works in this field are discussed in detail in this chapter.

The divide between the four levels is not sharp, and a full theory of science funding should account for all levels using a mixture of all the methods listed above, and possibly more. Such a project is well beyond the scope of this thesis. One of the main benefits of Kitcher's two recent books is their demonstration of how large this field of investigation is, and how little we know about it.[1] Nonetheless, we can make some progress by tackling each of these levels one at a time, producing insights that can provide tools or constraints for future work.

This thesis focuses on the level of choice between a set of individual projects (3b above), using tools from epistemology and theories of scientific methodology. The purpose of this thesis is to offer an account that will enable the evaluation of the rationality of current practices for choosing amongst project proposals (as discussed in Chapter 1), and that can be used in the design of more

[1] Although the main focus of Kitcher (2001, 2011), as far as the *choice* of research projects is concerned, is the most coarse-grained level, he touches briefly on each of the other levels: privatisation of research is discussed in Kitcher (2011, pp. 126-7), the motivations of scientists are discussed on p. 196, our ability to foresee the course of science is discussed on p. 120.

rational funding mechanisms (see example of such design work in Chapter 6). It is good to pause here and ask whether the project undertaken by this thesis is descriptive or normative. This thesis develops an account that serves both institutional critique and institutional design, and therefore the project has both descriptive and normative aspects. As described in the previous chapter (§1.4) the institutions of science funding serve multiple stakeholders, who sometimes hold contradicting (normative) preferences for the functioning of research funding allocation mechanisms. Given this setting, it is important for the account developed to contain descriptive aspects both of the features of funding allocation that are being evaluated (direct description), and of the normative preferences held by various stakeholders (description of normative preferences). In addition, the thesis develops a sceptical argument about the rationality of a general type of allocation mechanisms (prospective evaluation), an argument which combines descriptive elements, i.e. the presentation of reasons and evidence for the likely failure of such mechanisms in achieving their purported goals, with a normative recommendation, that in many cases such mechanisms should be replaced by other mechanisms that avoid this kind of failure. All of this is done while trying to remain as agnostic as possible about the specific aims science funding should achieve (normative agnosticism) as described in the next chapter.

The following section argues that a model-based strategy is useful for developing an account of research project choice in the context of funding allocation, relying both on features of the domain of investigation and general considerations of the usefulness of models for institutional design. The remainder of this chapter presents a survey of models developed by philosophers of science to address the topic of project choice.

2.2 The use of models for designing a science funding mechanism

Science, and science funding, are social institutions. As such, they are amenable to study by the methods of the social sciences. Two leading methods that offer themselves in developing an account of project choice in the context of science funding are model-building and ethnographic research, discussed below. Other methods are also relevant in this context, for example the qualitative survey of Chubin and Hackett mentioned in the previous chapter and more recent quantitative studies, including a significant component of statistical analysis, discussed in the last chapter. The discussion of modelling and ethnography below is meant as an accessible proxy to a more general comparison between theory-focused methods and evidence-focused methods, while side-stepping a discussion of the complex ways in which theory and evidence interact in these various methods.

Models Following Teller (2001), a model can be anything that is actively used to represent (some aspects of) a target system of interest by virtue of *similarity*.[2] The relevant aspects of the target system and the relevant criteria for similarity are determined by the model user's interests, and may differ from case to case. In our case, models of science funding should represent the desiderata for a science funding mechanism (see §1.4), in order of

[2]I take Teller's view to represent a consensus position regarding the philosophy of models in science. Teller presents a refinement of the model views developed by Giere (1988) and Cartwright (1983). The origin of a non-linguistic understanding of models can be traced to the semantic approach (Suppe, 1989), though Teller differentiates himself from this tradition by having models apply by *similarity* rather than by *isomorphism*. For a recent survey of the philosophy of models in science see Frigg and Hartmann (2012).

importance (e.g, representation of effectiveness will be more important than representation of fairness). We are free to choose any construct for use as a model, e.g. a mathematical game-theoretical model as in economics or a computational fitness landscape model as in population biology, as long as the relevance and similarity to the phenomena of interest are clearly stated. Models can be both descriptive and normative. The descriptive aspect of models is present in the aspects of the model that relate to aspects of the target system, though often models only describe a small subset of the rich phenomena in any given social context. A normative aspect can be added to models by introducing an evaluation function (e.g. a measure of utility to a game theoretical model) and comparing different configurations of the model using this evaluation function.

Ethnography Unlike model-building, which, on the whole, aims to make generalisations about the domain of interest, an ethnographic approach focuses on the detailed study of individual cases. Ethnography is committed to the "first-hand experience and exploration of a particular social or cultural setting on the basis of [...] participant observation. [...] It is this sense of social exploration and protracted investigation that gives ethnography its [...] character" (Atkinson et al., 2001, p. 5). The focus on a specific setting and the engagement in a protracted first-hand experience set ethnography as a very different approach from model-building. Like models, ethnography too has both descriptive and normative aspects. Clearly, the descriptive aspects of an ethnographic account are present in the rich account of the researcher's first-hand experience, and are often much more detailed than comparable descriptions found in models of the same domain. Ethnography can also serve a normative function; while it is debatable whether ethnographic accounts should include normative judgements based on the researcher's own values and culture, many ethnographic accounts present descriptions of the normative judgements of the participants in the culture being explored, and these normative judgements can carry weight in the design or critique of social institutions.

Of course, the choice of method need not be exclusive: there can be significant epistemic gain from mixing models and ethnography, as well as other methods of social investigation. However, we need to start somewhere. This section presents the argument for starting the investigation of science funding mechanisms using a model-based approach. The next section presents *domain-specific* reasons for preferring models over ethnography in the study of science funding. The section following presents *goal-specific* reasons for adopting models for the purpose of institutional design. It is important to repeat, though, that these are matters of degree, and ethnographic research is likely to complement and improve any modelling work carried out in this and future accounts of science funding.

2.2.1 Practical difficulties in studying decision-making in science funding

One feature that sets ethnography apart from models is its generation of novel empirical data. In ethnography, a rich empirical account is provided for a single setting. In model-building, on the other hand, a significant part of the process may take place with relatively little input from novel empirical data, e.g. in the explicit formulation and formalisation of background knowledge

and the investigation of relations between the model parts.[3] For a domain where the collection of novel data is readily accessible, we may prefer ethnography. However, at the time of writing this thesis, science funding is a domain where access to data collection is limited, at least with respect to the question being investigated. The reasons for this state of affairs are as follows:

Institutional secrecy The proposals submitted to peer review, and their merit evaluations, are almost always kept secret by funding bodies. While some funding bodies publish the funded proposals (albeit in short format), there is almost no funding body that allows access to its unfunded proposals. Furthermore, funding bodies do not allow access to the full process of its decision-making: while a few funding bodies have released anonymised sets of reviewers' scores (see Chapter 6), no funding body released verbatim reviewers' assessments, nor have protocols of review panels been released.

Problems of value judgements Making empirical evaluations of the quality of decision making in science funding is complicated by various problems of interpretation. First, the problem of internal merit evaluation (i.e. merit purely in virtue of contribution to the scientific enterprise) relies on a significant amount of tacit knowledge gained by socialisation in the relevant scientific discipline (see §1.5). Second, the problem of external merit evaluation (i.e. the value that a society would assign to a particular project or result) relies both on accepting an appropriate method of value aggregation, and on gaining access to the different values being aggregated. While models may abstract away or adopt an agnostic positions regarding specific value judgements (as described below and in the next chapter), such avoidance of value judgements is much harder in the immersive experience of ethnography.

Suppressed alternatives In an empirical evaluation of science funding we would like to make comparisons, both within a funding mechanism and between funding mechanisms. Within a funding mechanism, we would want to compare the funded and unfunded projects, for the purpose of estimating errors (see §1.6). Between funding mechanisms we would like to make comparisons of how well they meet the different desiderata. Unfortunately, such comparisons are not easy to make: we cannot make an *ex post* evaluation of unfunded proposals, because the funding is a necessary condition for the research to take place, and we cannot make comparisons between alternative funding mechanisms when the domain is dominated by a single mechanism (see Chapter 1).

For the above domain-specific reasons, it would be difficult to begin the study of science funding with methods that rely on novel empirical data. In addition to those difficulties, however, there are also positive reasons to opt for a model based strategy in the context of institutional design, as discussed below.

2.2.2 Usefulness of models for institutional design

Ultimately, the aim of this thesis is to offer an alternative science funding mechanism. As such, we can classify the work as "institutional design". The issue of institutional design, and particularly

[3] Of course, strictly speaking, we cannot learn anything new from creating a model that involves no novel empirical data, as it can take us no further than our already held background knowledge. Nonetheless, non-empirical model-building can prove useful as part of a larger, empirically-oriented strategy, as discussed later in the section.

the design of *economic* institutions, has been the centre of recent literature in the philosophy of social science. First, there is no question that institutional design work is being done by social scientists (chiefly economists). Furthermore, this kind of work has produced some remarkable successes, such as the design of a new clearinghouse for early-career medical positions in the US (Roth, 2002) or the auction of frequency bandwidths in the electromagnetic spectrum (Roth, 2002; Alexandrova and Northcott, 2009). Finally, the successes mentioned have made significant, though not exclusive, use of models. Both Roth and Alexandrova and Northcott offer insights into the benefits models offer to institutional design.

Roth

Roth (2002) introduces the notion of "design economics". In design economics, economists are asked not only to explain and predict features of existing institutions, but to design, partially or wholly, new markets and other economic mechanisms. Roth notes that the (predominantly rational-choice) theories and models, which feature prominently in the "old" economics, also feature in design economics, though in new roles. Traditionally, models in economics are used to highlight what are considered prominent and relatively stable generalisations, e.g. existence of optimal solutions, or the superiority of some algorithms over others in basic cases. In design economics, models serve the following purposes:

"Intuition" When coming to address a need for a novel institutional design, the variety of options can be daunting (in fact, if restrictions are sufficiently lax, the space of design options could be infinite). By going over the basic models offered by current theory for the domain of interest, the institutional designer can focus on the model, or set of models, that are closest to the phenomena at hand. Since models are used via judgements of similarity, a practised model user would be proficient at making similarity judgements quickly and carefully, and would therefore rapidly reach the set of models that are most suitable for the case at hand. These models serve as starting points, or intuitions, for exploring the design space, offering a heuristic for approaching the task.

Empirical tests Once the set of relevant models have been ascertained, a judgement of difference is made between the models and the situation facing the designer. If the situation is identical to a model case, a design can be offered immediately. However, this is almost never the case. Real situations offer many complexities, most of which would not already be represented in the available models. For example, the models may suggest a general class of possible design solutions, but fail to guide to one particular setup from within this class; alternatively, the models may suggest a single solution, but fail to give an accurate enough prediction of its expected outcome, a feature which may be required by decision makers. By analysing the differences between the relevant models and the case at hand, the designer is guided to search for relevant empirical data that would address this difference. If the data are already available, they may be incorporated into a novel model that would be more similar to the design case. Usually, however, such data would not be available. Thus, the models help identify lacunae in empirical data. These lacunae could then be addressed by direct empirical methods, e.g. laboratory simulations, where the design of the empirical investigation is informed by the model, and its results could potentially be incorporated back into the model.

Computer simulations While some empirical lacunae, identified using the available models, may be addressed using empirical methods, other unknown aspects of the target system may be too time- or resource-consuming to be tested in the lab. In such cases computational simulations of a wide range of alternatives could offer some help. While the exact epistemic merit of such simulations is debatable (Odenbaugh and Alexandrova, 2011), simulations can at least help us explicitly formulate, formalise and organise our background information about the target domain, in cases where the implications of background information are difficult to work out in one's mind. If we consider the space of design alternatives, computer simulations allow us to explore a cloud of possibilities around the model we are simulating. This may help us see that our background information extends further than we anticipated, or suggest that our models are more sensitive to some parameters than others. This in turn can help guide new empirical tests.

Presentation Institutional design often has a significant political aspect. A call for a new design often comes because some important stakeholder is unhappy with the current design, and until their trust can be regained for a new design, the operation of the institution is at risk. Thus, it is important for the institutional designer to be able to communicate efficiently and clearly their proposed design and its advantages to the various stakeholders. Models serve this purpose well, as they tend to be simple and easy to represent.

All of the above points are relevant in the context of this thesis. The *"intuitions"* provided by models of project-choice can help us look further afield than the currently dominant funding mechanism. *Computer simulations*, presented in Chapter 5, help visualise non-trivial ramifications of the available background information regarding science funding. Issues of *empirical data* are raised in various points along the thesis; in some cases they are met, using historical examples (Chapter 4) or statistical studies (Chapter 6), but overall the thesis leaves several empirical questions unanswered, motivating future empirical research. The issue of *presentation* has already been flagged as important in the context of science funding (§1.4), both in convincing society and its representatives to support science funding, and in convincing practising scientists to participate in a funding mechanism; one of the aims of this thesis is in serving as part of the presentational arsenal of those wishing to improve the system of science funding.

Alexandrova and Northcott

Alexandrova and Northcott (2009) offer a philosophical account of models in the context of institutional design. Like Roth, they agree that the usefulness of models is in an activity that bears similarity to engineering, or design. To account for that functionality, they develop a novel view of models as *open formulae*.[4]

Alexandrova and Northcott compare various philosophical accounts of models to the experience of institutional design of spectrum auctions. In the design of these auctions, theorists relied on and developed game-theoretic models of auction behaviour, while experimentalists provided novel empirical data that could inform the eventual auction design. Alexandrova and Northcott investigate the relationship between the theorists' models, the experimentalists' test beds, and the designers' proposals for auction rules. They convincingly argue that the models were not used to *derive* hypotheses for the expected auction behaviour, but rather that experimentalists

[4]The view of models as open formulae was also published as a separate paper by Alexandrova (2008).

and designers used the causal claims in the models while having the freedom to pick and choose from the models' assumptions, such that a causal claim may hold in the experimental case or the designed setup even if those settings did not satisfy the models' assumptions. For example, many models in auction game theory make causal claims about auction behaviour under strict assumptions about the distributions of participants' value assignments. However, in the required auction design the companies' value assignments were kept secret, and so the designers had to seek other conditions, different from those assumed by the models, that may bring about the causal relations between auction rules and auction behaviour specified by the models. The models did not, in general, offer guidance as to what these other conditions might be, but various conditions could be explored using the experimental test beds. This kind of pick-and-choose relation led Alexandrova and Northcott to offer an open-formula account of models.

According to the open-formula account, a model names, sketches or concretises a causal relation that *may* exist in the world, without specifying, or only partially specifying, the type of situations in which this causal relation occurs. Using the model one can generate a *hypothesis*, by claiming that in actual situations of a particular type, which are characterised to a level sufficient for picking out real situations, the causal relation *does* obtain. To clarify, the "open" nature is due to certain aspects of the target system which remain *unspecified* by the model, or even contradict some of the model assumptions, and these aspects are only "filled in" when the model is used in the generation of a hypothesis, i.e. when a claim is made about the model's causal relation applying to a real situation.

It is this *open* nature of models that facilitates their use as tools for design of social institutions. There is a myriad of causal relations that influence the operation of social institutions (and other complex systems), so that in general, most possible configurations of the institutions have limited predictability. However, the use of models, alongside empirical data, can help us zoom in on the few configurations that are tractable, and from them pick a configuration for which we predict beneficial outcomes. Put more succinctly, models, alongside empirical data, help us design institutions that operate in the way we want, within the given set of constraints.

To conclude, we've seen that models *are* used for institutional design, and that there are *good reasons* for using models for institutional design. Given the understanding of models as open formulae, I argue below that models are also suitable for sceptical work, i.e. the prediction that some institutional designs would be less optimal than others or fail to meet the design criteria.

2.2.3 Usefulness of models for sceptical work

While this thesis argues for a novel institution of science funding, it also presents a critique of peer review, the currently dominant funding mechanism (see Chapter 1). As such, the work undertaken is simultaneously institutional design (of the alternative) and scepticism (regarding the effectiveness of peer review). The same features that allow models to be useful in institutional design are the ones that make them effective for sceptical work.

Sceptical arguments based on models can be used to criticise current or proposed institutional designs, or place restrictions on the kinds of configurations that are considered in the process of institution design. For example, the Laffer Curve in economics is a simple model of the effects of tax rates on tax revenues (Laffer, 2004). According to the model, tax rates have two causal effects on tax revenues: an "arithmetic" effect, which increases revenue as tax rates increase, and an "economic" effect, which lowers revenues at high tax rates as agents in the economy reduce

production or seek tax-avoidance mechanisms. While the Laffer Curve cannot be used for direct design of tax rates, as the behaviour in mid-range tax rates is poorly described by the model, it can be used as a sceptical model that rejects very low (near 0%) or very high (near 100%) tax rates if the goal is to maximise tax revenues.

As will be fleshed out in the following chapters, the sceptical argument of this thesis roughly follows the following steps:

1. The rationale for peer review is that peer assessment raises the overall merit of projects selected for funding.

2. In the model developed in the thesis, successful peer assessment relies on an ability to predict future merit evaluations, and to estimate merit for hitherto unexplored approaches.

3. Empirical data, generalised using the model, shows peers are unlikely to be able to make such judgements in a reliable manner, at least in an important subset of cases.

2.2.4 The models in this chapter

The above should make it clear why a model-building strategy is adopted in this thesis. However, the construction of a model that can be directly applied to the context of science funding only begins in the next chapter. The existing models considered and criticised in this chapter were not originally presented by their authors with the aim of furthering the institutional design of science funding mechanisms (except the model presented by Peirce (1879/1967), which does address this problem directly). Nonetheless, they are useful for the work carried out in the remainder of the thesis for these reasons:

Incremental development Model building is a difficult and time consuming activity. It is therefore more efficient to develop existing models, or at least to borrow conceptual, mathematical, computational and representational tools from them, than it is to start from fresh.

Explication of assumptions A central argument of this thesis is that evaluations of scientific merit are difficult, and in certain cases impossible, to make. In the general literature on science funding there has been little treatment of what is involved in evaluations of scientific merit, or even what scientific merit *is*.[5] One place where such issues are discussed explicitly is in the literature of models regarding project choice, discussed in this chapter. This is not surprising, as model-building often requires the explicit formulation of background knowledge and assumptions. As the models discussed were developed by philosophers of science, it is not unreasonable to treat the assumptions made in these models as approximations of informed attitudes towards scientific merit and its evaluation. When these attitudes are found to be faulty, their existence in the models suggests the existence of widespread misconceptions regarding scientific merit, which in turn increases the relevance of the arguments presented in the thesis.

The remainder of this chapter is dedicated to discussion and criticism of two generations of models of project choice, following the methodological guidelines set out above.

[5] Polanyi's paper discussed in §1.5 presents an important exception, which is why it was chosen as the strongest defence of peer review.

2.3 Looking for Kitcher's lost dog: Research project choice as utility maximisation among finite alternatives

An important strand in philosophy of science dealing with the choice of research projects is the work done on "division of cognitive labour". Division of cognitive labour deals with the assignment of the available cognitive workforce, the scientists, to possible research projects. There have been three important works in this strand:

1. Peirce (1879/1967), "Note on the Theory of the Economy of Research",
2. Kitcher (1990), "The Division of Cognitive Labour", which was later developed and expanded into Kitcher (1993), *The Advancement of Science*, ch. 8: "The Organisation of Cognitive Labour",
3. Strevens (2003), "The Role of the Priority Rule in Science".[6]

All three works feature prominent philosophers of science utilising mathematical economics-inspired models to tackle the question of the optimal choice of research projects, or rather the optimal assignment of investigators to research projects. The arguments supported by the models are different in each of the works, as described below. However, the focus is less on the specific arguments advanced in each of the works, but rather on the similarity of approach and models, which form the flawed but promising precursors of the tools deployed in this thesis.

The following presents quick, stand-alone overviews of each of the works. The three overviews are followed by a detailed discussion that considers all three together.

2.3.1 Introducing the arguments

Peirce

By far the earliest of the works is a paper by the famous pragmatist philosopher Charles Sanders Peirce. In the paper, Peirce provides a framework for the optimal allocation of resources to research projects ("researches" in his terminology). Peirce starts from the assumption that quantitative research projects aim at reducing the "probable error" (in his terminology) of a certain quantity, and that with some extension of the concept (which he does not provide), the same could be said for qualitative research projects as well.

Peirce then associates the reduction in probable error of a specific quantity with a certain amount of utility. The resulting trend of utility as a function of uncertainty reduction must display, Peirce argues, diminishing returns. In parallel, Peirce associates the reduction in uncertainty with a cost. The resulting trend of payoff (decrease of probable error) as a function of cumulative investment also displays, according to Peirce, diminishing returns. Peirce supports the existence of diminishing returns for both utility and cost with reference to the history of science and normal human experience, in the following rather morbid quote:

[6] Another work on roughly the same topic is Brock and Durlauf (1999). This field has many similarities, but also some important differences, with the field of the economics of scientific research, which includes seminal works such as Nelson (1959), Arrow (1962), Dasgupta and David (1994), and Diamond (1996). The latter field seems to have been superseded in the last decade by the study of the economics of innovation. Still, many of the foundational questions, raised early in the history of the literature described above, are worth revisiting today, as argued in the remainder of this thesis.

> We thus see that when an investigation is commenced, after the initial expenses are once paid, at little cost we improve our knowledge, and improvement then is especially valuable; but as the investigation goes on, additions to our knowledge cost more and more, and, at the same time, are of less and less worth. Thus, when chemistry sprang into being, Dr. Wollaston, with a few test tubes and phials on a tea-tray, was able to make new discoveries of the greatest moment. In our day, a thousand chemists, with the most elaborate appliances, are not able to reach results which are comparable in interest with those early ones. All the sciences exhibit the same phenomenon, and so does the course of life. At first we learn very easily, and the interest of experience is very great; but it becomes harder and harder, and less and less worth while, until we are glad to sleep in death. (p. 644)

Peirce then presents mathematical equations which define the relations between utility, cost, and the decrease in probable error. By equating the marginal utility per cost across multiple research projects (equilibrium condition), he provides a prescription of the optimal allocation of available finite resource across a range of possible research projects with known utility and cost functions. For a detailed discussion of Peirce's paper, see Wible (1994).

Kitcher

The most influential of the three works mentioned is Kitcher (1990), which was later expanded and reworked into the final chapter of Kitcher (1993). There are several related arguments advanced by Kitcher, using the mathematical economic-inspired approach, especially in Kitcher (1993). However, the overarching argument, which is a continuation of the overall theme of the book albeit in a different style, is a defence of the rationality of science in a response to works in the sociology of science which are seen as potentially undermining this image of rationality. The details are not crucial here, but an overview of the argument is roughly as follows.

I propose the following formulation as a summary of Kitcher's version of "the argument from the sociology of science":

P1 We have an articulated image of science as the rational pursuit of (scientific) truth.

P2 When we cash out this rationality in the context of scientific activity, we get a prescription of behaviour that would most likely lead to an increase in truth.

P3 When we observe the actions of actual scientists, their actions are different from those prescribed by scientific rationality. Specifically, their actions reflect a motivation towards individual rewards and prestige, and not a disinterested pursuit of truth.

C1 Either scientists are not rational, or the rationality of scientists is not directed towards the pursuit of truth.

Kitcher accepts premises P1-P3, but rejects the conclusion C1. His argument relies on his claim that there is an important distinction to draw between the actions of individual scientists and the products of the scientific community as a whole. I propose the following as a summary of his response:

P4 There can be several different approaches to resolve a particular valuable scientific problem.

P5 Approaches differ in terms of their reliability, and in their probability of success given the number of scientists pursuing them.

P6 The community optimum for maximising the likelihood of resolving a problem often lies in distributing the available scientific workforce across several approaches, rather than assigning all of them to the likeliest approach. This means that for truth generation to be maximised, some scientists should be assigned to projects the scientific consensus considers less likely to generate truth.

P7 A reward system can be designed so that scientists who are motivated by personal rewards ("sullied scientists" in Kitcher's terminology) will allocate themselves in a way that reflects the optimal distribution from the community's perspective. Specifically, their distribution will perform better than the distribution that would result if each individual scientist would allocate themselves to the project they consider most likely to generate truth, even when accounting for natural difference of opinion in the community.

C2 When the pursuit of truth is considered at the community level, pursuit of rewards and prestige at the level of the individual scientist does not necessarily undermine, and in fact can help achieve, optimal pursuit of truth. Thus, the observed deviation of individual scientists from the likeliest methods due to the pursuit of personal gains is not proof of the irrationality of science and does not support the claim that science does not aim at truth, as long as we consider rationality and the pursuit of truth at the level of the community.[7]

It is interesting to note that (Longino, 2002) draws a connection between Kitcher's models and the decisions made by science funding bodies, though she does not pursue it:

> In spite of the exuberance of calculation, it is still a little difficult to understand how this apparatus should be deployed. [...] One possibility [...] is that there is a centralized decision-making system. But so far I know this exists only in Bacon's New Atlantis. The closest approximation is *national funding bodies*, and they depend on proposals for research coming up from the "grass roots" (and in the United States are themselves multiple and decentralized). There is no doubt that funding bodies engage in some kind of cost-benefit calculations in making funding decisions. It would be interesting to know what goals they try to optimize and whether their decision making is adequately modeled by Kitcher's formulas. (Longino, 2002, p. 73, my emphasis)

Later chapters show how Kitcher's models of project choice, and the other models discussed in this chapter, can form the basis for models that can be used to evaluate science funding decisions, though in their present form they are inadequate for the task, as shown later in this chapter.

[7] For a general review of the argument in Kitcher (1993) see Lipton (1994). The specific argument about the division of cognitive labour is discussed by Goldman (1999, pp. 255-258), who juxtaposes it with a discussion of funding decisions by the NIH (pp. 259-260). For a detailed criticism of Kitcher's general argument see Longino (2002, pp. 51-68). Longino follows up, in an appendix (pp. 69-76), with a criticism of the "division of cognitive labour" arguments.

Strevens

Strevens (2003) presents an explanation for the social phenomenon of the "priority rule" in science, the assignment of all rewards to the first scientist, or group of scientists, who make a certain discovery, as opposed to a more equal distribution of rewards between the first discoverers and the runners-up. Strevens claims that his explanation is more successful than the norm-based explanations provided before him, e.g. by the famous American sociologist of science Robert Merton.

Strevens claims that the "priority rule" is an extension of a familiar social reward mechanism, one that rationally maximises the production of a certain social good. The reason that the priority rule scheme is not widely adopted in other social contexts (e.g. all credit going to the first teacher whose students achieved particularly high grades) lies not in a particular choice of reward structure in science, but rather in the structure of the good itself, or rather in the scheme of social utility generated from the good. In the case of scientific facts, Strevens argues, the first successful discovery confers virtually all public utility, whereas the second and third discovery of the same fact yields virtually no public utility. That this is the actual structure of public utility for scientific discoveries is, according to Strevens, "a consequence of the very nature of scientific information", and as a consequence "additional discoveries of the same facts or procedures are pointless" (p. 1).

In the bulk of the paper, Strevens establishes the link between this "very nature of scientific information" and the "priority rule". He does so using a mathematical economic-inspired framework inspired by Peirce (1879/1967) and developed from Kitcher (1990). Strevens outlines, and then formalises, three possible rewards schemes: "Marge", which rewards scientists according to their (*ex ante*) expected marginal contribution to the success of the research project they join; "Goal", which rewards scientists according to their (*ex post*) actual marginal contribution to the success of the project; and "Priority", which is similar to "Goal" except only the scientists who took part in the first project to succeed are rewarded. Using a mathematical model of the public utility of research products, where utility decreases to zero once the first success has been obtained, Strevens shows that "Priority" outperforms both "Goal" and "Marge".

2.3.2 Shared conceptual framework

At this stage of the thesis, the details of the arguments offered by these three philosophers are less important; the important aspect is the modelling framework shared by these three authors in their formulation of their arguments. As discussed below, this framework entails some substantive assumptions about the nature of research, and it is an important step of the thesis to challenge these assumptions.

The most readily noticeable commonality between the three works is the presentation of arguments and models in the form of equations. Here is one example from each work:

> [...] The general answer is that we should study that problem for which the economic urgency, or the ratio of the utility to the cost
>
> $$Ur \cdot dr/Vs \cdot ds = r^2(Ur/Vs) = (h/k)(r^4/\sqrt{a + r^2}$$
>
> is a maximum. (Peirce, 1879/1967, p. 645)

[...] Division of labor is thus preferable if there's an n, meeting the "give both a chance" constraint, $m < n < N - m$, such that

$$q_1 p_1 * (n) u_2 + q_2 p_2 * (N - n) u_2 > q_1 u_1 - q_2 u_1$$

(Kitcher, 1993, p. 355)

[...] Then the optimal value of n will be the value for which

$$m_1(n)(1 - s_2(N - n)) = m_2(N - n)(1 - s_1(n)).$$

(Strevens, 2003, p. 19)

The meaning of these particular equations is not important, as they are but a small selection of many. The important lesson here is that the equations are meant to set up the problem of the allocation of resources/investigators to projects as an *optimisation problem*. The technique is an application of rational choice theory, which is frequently used in microeconomic modelling; it creates a parameterised model of the decision the agent (the scientist) is facing. It then extrapolates the community outcome, assuming each individual will choose the best option presented to them. This community outcome is presented as a function of the parameters that influence the individual decisions of the agents. Once the equations are set up, their form directly suggests the existence of an "optimal" configuration, in which the expected utility of the community is maximised, to which all other configurations can be compared. The "optimality" of the optimal configuration is known to anyone who knows the values of the parameters involved in the equations, and it is therefore an "objective" or "rational" optimum, to be discerned for example by a "philosopher monarch" (Kitcher, 1990, p. 8).

Not only the presentation, but also the theoretical concepts used, are borrowed from economics. The parameters in the equations relate to the "utility" accrued from a (successful) research project, and the "cost" or "labour" required to achieve success and gain this "utility", where likelihood of success is represented as a function of cost/labour. Like other goods, research labour and research output display "diminishing returns". The motivation and justification given for external intervention, e.g. by funding or award bodies, in the organisation of research, is that the products of research are described as "social goods", and therefore warrant intervention by external players (non-scientists), who represent the society that stands to benefit from these "goods".[8]

All three authors mention, at least in passing, the tension, either real or commonly perceived, between a search for (objectively valuable, socially beneficial) truth and a search for (subjectively valuable, personally beneficial) rewards. All three authors state their belief that the "economic" approach can shed light on this tension. Peirce is dismissive of accounts that focus on optimising for researchers' personal gain, stressing that the framework he develops applies when "the object of investigation is the ascertainment of truth" (Peirce, 1879/1967, p. 678). Kitcher and Strevens, in contrast, believe that the best way to gain socially beneficial, objective truth is by incentivising scientists, appealing to their desire for subjectively valuable, personal benefits. Strevens uses the economic framework to link the social good that accrues from (assumedly true) scientific

[8] For a further survey of the "economics approach" in the philosophy of science, which links the three works mentioned above and situates them in a wider context of the philosophy of science, see Strevens (2011).

information to the reward structure of science, and uses that link to explain the priority rule (Strevens, 2003, p. 32). Kitcher, in addition to explaining existing reward structures, urges us to apply the analysis more critically, a practice which may lead us to a redesign of the existing mechanisms (Kitcher, 1990, p.22).

This thesis follows Kitcher's advice stated above, and considers how philosophy of science can be used to change and improve the institutions of science. However, in order for improvements to be considered, the failings of current institutions need to be highlighted. This thesis argues that certain decisions made within scientific institutions are guided by misconceptions about science. These misconceptions can be revealed by exploring the models of project choice discussed in this section, in which they appear as idealisations. I therefore explore in greater detail the idealising assumptions shared by these models.

2.3.3 Substantive assumptions of the shared framework

The three accounts mentioned above contain substantive assumptions about the decisions made by scientists: assumptions about what scientists know, about what scientists want, and about how they choose. The list below highlights assumptions that are entailed by the assignment of a *single*, *known*, *universal* (among scientists) utility value to each potential research project, and a *single*, *known*, *universal* probability of success (in the form of a success function) to each potential research project. Thus, on this formulation, the substantive idealising assumptions are:

Clear goal Potential research projects are defined by their goal, which can be clearly expressed, *ex ante*.

Known benefit The clearly defined goal of each potential research project has a utility or utility function which is known by the relevant decision makers (scientists themselves in Kitcher's and Strevens' models, a resource allocator in Peirce's model). The known utilities allow the decision makers to make cardinal comparison between projects.

Finite set of alternative research projects At any given point in time the decision maker is faced with a finite set of well-defined potential research projects.

Predictable reliability of method/likelihood of success The likelihood of success of a potential research project, which specifies a specific method for attempting to obtain a defined goal, is known in advance, either as a representation of the reliability of the method or as a function of the amount of resources/workforce invested in pursuing it.

The image of science that emerges from these substantive assumptions is clearly conveyed by Kitcher's story of looking for the family dog that ran away:

> When we lived in California, Bertie, the much-loved family dog, would come and go freely between the house and the fenced backyard. One afternoon, two or three hours before sunset, when coyotes became active, someone came to read the meter and left the gate to the yard open. When Bertie went out, he spotted the chance of adventure and set off to explore. Discovering his absence we had to formulate a plan for finding him (quickly, since he's too small to take on coyotes). We normally walked him along one of two routes, which we assumed were the places he'd be most likely to go. One

route was estimated to be a more probable path than the other. There were four of us. How to proceed? (Kitcher, 2001, p. 111)[9]

The lost dog story embodies all four substantive assumptions: the *clear goal* is to find the dog before sunset, and all interested parties (the family members) know what the dog looks like and what would count as finding it before sunset; the *known benefit* is that finding the dog before sunset would prevent harm coming to the dog, which all interested parties agree is a good and important goal to pursue; the *finite set of alternatives* is represented in searching in one of the two routes the dog was usually walked along; and the *predictable likelihood of success* is captured by the estimation that the dog would probably be along one path rather than the other.

The next section presents a criticism of this depiction of science, arguing that it is at best useful for only a small range of real scientific situations. This view needs to be replaced, via modification of the basic assumptions, if we want to have a working model that can be used to analyse science funding decisions.

2.3.4 Criticism of the assumptions

The general line of criticism presented below, combining comments from other scholars with my own criticism, is similar to the line of criticism against most modelling work: the assumptions are idealisations which fail to correspond to reality, either in most cases (idealisations fail to capture normal cases) or in a range of important cases (idealisations fail to capture significant cases). Furthermore, when these idealisations are replaced with more accurate representations of reality, the models change to the extent that they produce results which no longer agree with the results arrived at under the idealised assumptions. In other words, the models lack fidelity.

The criticism of lack of fidelity does not necessarily undermine the arguments presented in the works mentioned above. Peirce, Kitcher, and Strevens can all counter that, *for the cases significant for their arguments*, the idealisations hold. Especially in the case of Kitcher's argument, the defended position is quite humble: lack of truth-maximising behaviour in individual scientists does not necessarily imply lack of truth-maximisation at the community level; since Kitcher only needs to sketch a single, somewhat plausible scenario in which truth-maximisation diverges between the individual and community levels, his models serve his purpose well, and are not undermined by the existence of idealisations in his assumptions. However, this thesis goes further than this humble conclusion, and builds upon models of research project choice to analyse the decision making of funding bodies. For such extension work, the fidelity charge needs to be taken seriously.

The limited extension of the substantive assumptions in the economics-based models has been commented on by Kitcher (2001) himself, as well as by Muldoon and Weisberg (2011). These criticisms are presented below, concluding with a summary and some notes of my own.

Kitcher's criticism

Already in *Science, Truth and Democracy*, where the analogy was first introduced, Kitcher points out the limitations of the lost dog analogy when thinking about science. I propose the

[9]The reference to the parable of the lost dog is not a single incident; Kitcher returns to this image of the search for a lost dog in his most recent book (Kitcher, 2011, pp. 193-4).

following paraphrase of Kitcher's criticism, according to which the reasons that undermine the generalisation from the lost dog analogy to all scientific research are:

Epistemic significance is dynamic and subjective *If* there is no context-independent notion of significance,[10] or "objective" significance, then any evaluation of significance will depend on contingent ideas about what is significant, and what counts as epistemically valuable, which are the quantities the reward structure is meant to maximise. As these are likely to evolve, sometimes drastically, during the course of investigation, we may find that what we thought was significant is no longer so, or vice versa, meaning that expected, *ex ante*, optimality, based on the model, will fail to correspond to actual, *ex post*, optimality, when assessed with hindsight. After two hours of looking for the dog, we may realise we actually just want ice-cream, and don't care about the dog so much.

Lack of robustness resulting from lack of knowledge Even if we had "objective" significance, the model, which was developed for the case of two competing approaches, may fail if we change the parameters. At present, we know very little about the reality we are trying to model: what are the choices of projects faced by scientists, and what reward structures they are facing. If this reality differs significantly from the one represented in the model, then the local optimisation arrived at by the model may well lead to less-than-optimal results in different, perhaps significant, cases. Perhaps most dogs, but not Kitcher's dog, return home after a while, and going out for a search, which was optimal for Kitcher's dog, is a waste of effort for most dogs.

Reward structures can have negative effects While the model shows how reward structures can have a positive effect on the division of cognitive labour, it ignores the potential negative effects of encouraging scientists to pursue public acclaim, such as pandering to pre-existing popular prejudice. In a hypothetical case, where the evidence could be equally interpreted as either challenging or supporting a widespread belief, reward-orientation may cause scientists to favour supporting the belief, as they are likely to face less public criticism or challenges and more easily reap the rewards. Other outcomes of reward-orientation, less likely but more extreme, are also possible, such as increased motivation to engage in over-hyping results or to commit fraud. Perhaps one young member of the family will get so excited about being congratulated for finding the dog, that he will secretly leave the door open in the future.

Kitcher does not present an alternative model that would address these criticisms. Rather, he simply states that these criticisms show that developing a general methodology of the organisation of investigation is *difficult*. We may want an optimal system, but we are very far from having the knowledge required to design one. Nonetheless, he claims, models of the system can help us avoid really bad configurations, and may occasionally show us how to resolve tricky situations (he gives examples of neither).

Kitcher recovers from the criticism by widening the scope of his analysis. He outlines three possible questions about investigation:

[10] Kitcher (2001) introduced the notion of "significance" to differentiate between valuable truth, or truth worth pursuing, and just any old truth. The notion is not present in the earlier model-heavy work, but "significant truth" can be identified, at least in the context of this criticism, with "high-utility truth", which is a central concept in the earlier work. A more complete discussion of measures of scientific merit is presented in the next chapter.

1. What are good policies for individuals to adopt if they want to learn epistemically significant truths?
2. What are good ways for communities to organize their efforts if they want to promote the collective acquisition of epistemically significant truth?
3. What are good ways for communities to organize inquiry if they want to promote their collective values (including, but normally not exhausted by, the acquisition of epistemically significant truth)? (Kitcher, 2001, p. 114, changed format of numbering for consistency)

Kitcher notes that the first question has been the focus of much philosophical work, but that it is very hard to provide any general context-independent answers for it. He then claims that the second question, for which the models developed by Peirce, Strevens, and himself seem to provide an answer, is also likely to fail to provide a full context-independent answer. He therefore chooses to focus on the third question, endeavouring to show that, despite it looking even harder than the first two, it is actually more tractable. Once Kitcher's focus has shifted, he no longer comments on the economics-inspired model, though he continues to refer to it in later work.

While Kitcher's own work seems to have moved away from the economics-inspired models, the field is far from neglected. Strevens' work, discussed above, is quite recent, as are Weisberg and Muldoon's papers, both the critical and the constructive, discussed below. Outside philosophy of science, similar models have been picked up and expanded on by the growing field of Scientometrics (Scharnhorst et al., 2011).

Using the four levels of course-graining introduced in §2.1, it is easy to account for the shift in Kitcher's position: Kitcher is shifting his scope, setting aside the epistemic, psychological, and sociological/economic levels, and focuses on the political level. Unfortunately, in doing so he leaves behind a flawed model at the epistemic level, but still continues to refer to it in later work, without addressing or even recounting the criticisms he himself noted. So far we have only seen some of the flaws in this type of model; the rest are presented in the following sections.

Muldoon and Weisberg's criticism

Muldoon and Weisberg (2011) directly criticise the fidelity of idealisations in Kitcher (1990, 1993) and Strevens (2003). They identify four key idealising assumptions shared by the models:

Rational agents Scientists are rational utility-maximisers.

Finite number of pre-defined projects Each scientist chooses to work on one of a finite list of possible projects, all of them known in advance to all scientists.

Known distribution Each scientist knows the decisions made by all other scientists, i.e. they have full knowledge of the distribution of scientists among projects.

Known success function All scientists agree on the success function of each project, which defines the likelihood of success as a function of the number of scientists pursuing the project.

It is worthwhile to compare Muldoon and Weisberg's list to the list presented in §2.3.3. While the overlap is significant, Muldoon and Weisberg fail to highlight the problematic assumption

that the benefit (either as utility or as significance) of a research project is known in advance and universally agreed upon. The terminology might be confusing, but the last item on their list, "known success function", overlaps not with "known benefit" but rather with "predictable likelihood of success". Thus, when they offer their remedial model, discussed in the next section, they also build into their model single-valued significance that is universally agreed upon and time-independent.

Of these four idealising assumptions, Muldoon and Weisberg claim that the latter two, which they label the *distribution assumption* and the *success function assumption*, are not robust: when they are replaced with more accurate representations of reality, as they do in their own model described below, the model yields different results. Specifically, the result these assumptions undermine is optimality; both Kitcher and Strevens claim that, given the reward structure they specify, optimal division of cognitive labour is obtained. Muldoon and Weisberg argue that, while all other elements of the model are kept the same, when either or both of these assumptions are replaced with more realistic representations, the suggested reward system no longer leads the scientific community to an optimal division of cognitive labour, according to the definition of optimal division set out be Kitcher and Strevens.

To support their claim, Muldoon and Weisberg make a transition from marginal contribution/reward (MCR) models to simulated agent-based models. The transition follows from the fact that the two idealising assumptions cannot be relaxed in MCR models, as they allow no variation in information and judgement between agents. The price that needs to be paid in the transition is that, unlike MCR models that lend themselves to closed-form mathematical analysis, agent-based models require computational simulation. To establish the validity of the transition, Muldoon and Weisberg first recreate the conditions stipulated by the two assumptions, by giving all agents in the model full information of all other scientists' decisions, and by assigning to all of them the same estimation of success functions. Under these conditions, Muldoon and Weisberg were able to replicate the results reported by Kitcher and Strevens.

Muldoon and Weisberg's criticism of the distribution assumption relies on the observation that, in most cases, scientists do not know what projects all other scientists work on, even within a small sub-discipline, because:

> [T]here are too many scientists, too many research programs, and too much physical distance between scientists to maintain the level of communication necessary for this knowledge. Scientists are much more likely to be informed of the work being carried out by colleagues in their own laboratories, colleagues in close proximity, and those with whom they have pre-established relationships. (Muldoon and Weisberg, 2011, p. 166)

Muldoon and Weisberg introduce a realistic alternative to the distribution assumption by representing the agents in the model as spatially separated, and by attributing to them a "radius of vision", a parameter in the model which defines how many other scientists each agent "sees", where to "see" a scientists is to have knowledge about their project choice. They show that when the radius is decreased, the resulting distribution of cognitive labour is no longer optimal. They consider the potential counter-criticism that scientists with limited vision would not just make decisions based on what they "see", but would rather extrapolate from what they "see" to the rest of the relevant scientific community. However, they reject this counter-criticism, by noting that

1. the local community is unlikely to be representative, as communicating scientists are more likely to work on similar problems, which will introduce a bias in the estimate, and

2. there are no available mechanisms by which scientists can estimate the size of the relevant scientific community.

Muldoon and Weisberg's criticism of the success function assumption begins with asking how one should interpret the probability of success of a research project. They offer three alternatives:

1. Frequentist: what portion of research groups that pursued this project in the past have succeeded?

2. Objectivist: what is the inherent potential of success in this project?

3. Subjectivist/Bayesian: what is the subjective probability of success each individual scientist assigns to the project when she makes her decision about project choice?

They reject the frequentist interpretation, on the grounds that research projects, both successful and failed, are rarely repeated. They reject the objectivist interpretation, on the grounds that the premise for the project is either true (as in the case of projects aiming to elucidate the structure of DNA), in which case the probability of success is one, or it is false (as in the case of phlogiston chemistry), in which case the probability of success is zero.[11] Thus, they are left with the subjectivist interpretation.

Under the subjectivist interpretation, the success function assumption, that all scientists assign the same probability of success to all projects, would rely on all scientists holding exactly the same posterior probabilities for all relevant information. Muldoon and Weisberg reject the possibility that a convergence of prior probabilities can account for this similarity in posteriors, and they express doubt that there is any mechanism that will enforce convergence of posteriors to a level of complete uniformity. Thus, the success function assumption also becomes related to communication flow between scientists, and the observation of imperfect and/or limited communication requires us to relax the assumption. Muldoon and Weisberg model this relaxation by allowing scientists' success function to vary as a normal distribution around the parameters defined by the model as the "true" value. When this relaxation is simulated, the resulting distribution no longer achieves optimality.

Possible responses to the criticisms

In both Kitcher's and Muldoon and Weisberg's criticisms, a major weakness of the model was identified in faulty assumptions about the knowledge individual scientists hold. In Kitcher's criticism, the fault was in assuming that all scientists had correct, unchanging knowledge about significance; Muldoon and Weisberg also raise this criticism (the success function assumption), and add to it a criticism of the assumption that all scientists know about the decisions of all other scientists.

Given such criticisms, a potential model user, such as a designer of a science funding mechanism, is left with several possible responses:

[11] I disagree with Muldoon and Weisberg on this point. However, this disagreement is tangential to the current presentation of their criticism. I will return to it in much greater detail in the next chapter, when discussing epistemic fitness.

Empirical evidence By providing empirical evidence that scientists in actual situations *do* have the relevant information, or at least a good approximation of it, one can reject the criticisms and use the models as they stand.

Change the scope It might be possible to identify contexts in which the assumptions about information *are* likely to hold. For example, individual scientists may not know what others in the community are doing, but a central institution, such as a national funding body, may have such information (it may in fact have great influence over the content of this information).

Change the models It is possible to deal with the criticisms by creating models that do not rely on these assumptions, i.e. models that have explicit representation of limited knowledge. The next section discusses the agent-based models developed by Weisberg and Muldoon, that attempt to address the criticism in this way.

Change reality If limited information is the problem, it might be possible to introduce new technologies, practices, and institutions, whose role will be to facilitate information transfer between scientists. If one succeeds in doing so, one can *make* the information assumptions robust.

Scepticism Finally, it might be the case that the required information simply cannot be had, under any circumstance and by any agent. If one can demonstrate this or at least strongly argue for this, it will place a clear constraint on the kinds of models that can be developed, and will also teach us something interesting about the nature of social investigation.

The next section considers Weisberg's and Muldoon's agent-based model. This model provides a constructive alternative to the MCR models criticised above, and thus follows the "change the models" option from the list above. However, as described below, Weisberg and Muldoon's model also contains unrealistic assumptions that make it unsuitable for designing a science funding mechanism. In the following chapters two of the approaches above are adopted. The next chapter adopts the "change the models" strategy by constructing a novel model for the domain of resource allocation for research. The chapters following rely on this model to develop a general sceptical argument about the information that can be had about research projects in advance, an argument that has direct bearing on the design of science funding mechanisms.

2.4 Mavericks and followers: Hill climbing in epistemic landscapes

Weisberg and Muldoon (2009) are explicitly appreciative of the work done by Kitcher and Strevens, which was discussed in the previous section. They believe these models show that a look at the micro-motives of scientists, and their relation to the macro-level distribution of cognitive labour, reveals a complex structure that defies the maxim that what is epistemically good for an individual is epistemically good for the community, or vice versa (Weisberg and Muldoon, 2009, pp. 226-7). "Epistemically good" in this context means "contributes to the generation of significant truth".

However, Weisberg and Muldoon move away from the models of marginal utility per scientist per project discussed in the previous section. Instead of looking at specific races for a known

outcome, such as finding a lost dog or finding the structure of a very important molecule, they look at the case of a group of investigators exploring a (potentially) fertile epistemic landscape. The shift to a landscape model enables the amelioration of the problematic assumptions they highlighted in Kitcher and Strevens' models: since agents are distributed around the landscape, it is possible to define a "vision" parameter, which limits the amount of knowledge each agent has about the actions of other agents, thus addressing the "distribution assumption"; since agents are individuated, it is possible to assign, simultaneously, different success functions to different agents, thus addressing the "success function assumption". As their analysis shows, such landscapes, when populated by agents who are individuated in such a manner, are best explored, both in terms of rapid discovery of significant results and in terms of maximal cover of the epistemic landscape, by a mixture of competition and cooperation, of "mavericks" and "followers". Note that this type of analysis, about the optimal *characters of researchers*, addresses a different loci of questions than the account developed in this thesis, which focuses on the optimal *choice of projects* (see §2.1). Nonetheless, Weisberg and Muldoon make a significant contribution in developing more realistic models that address the same target domain, and therefore their model merits a close examination.

Weisberg and Muldoon's "epistemic landscape" is defined using three components: a "topic", "approaches" within that topic, and "significance" assigned to each of the approaches. Each landscape corresponds to one, and only one, topic, and this topic defines the boundaries of the landscape. The scope of a "topic" can be broader or narrower, depending on the resolution required by the model-user. Weisberg and Muldoon focus on narrower topics, which "approximately correspond to the topic that a specialised research conference or advanced level monograph might be devoted to" (Weisberg and Muldoon, 2009, p. 228). By identifying each epistemic landscape with a single topic, Weisberg and Muldoon already limit the scope of applicability of their model: it cannot help us compare approaches if they belong to different topics. In the context of science funding there is often a need to compare projects from different topics, as the number of topics vastly exceeds the number of independent decision-making units within funding bodies. This limitation is addressed in later chapters.

Within a specific topic, "approaches" are:

> narrow specifications of how an individual scientist or research group investigates the topic. An approach includes:
>
> 1. the research question being investigated,
> 2. the instruments and techniques used to gather data,
> 3. the methods used to analyze the data, and
> 4. the background theories used to interpret the data. (Weisberg and Muldoon, 2009, pp. 228-9)[12]

Finally, Weisberg and Muldoon adopt Kitcher's notion of scientific significance, and assign to each coordinate in the landscape (i.e. to each approach) a height which represents the significance associated to that approach (see description in the next paragraph). They agree with Kitcher

[12]It is interesting to note that "approaches" are the same for all scientists in Weisberg and Muldoon's model, which means the reference to an "individual scientist" should be read as something like "the average scientist", "a model scientist", or "an ideal scientist". These readings differ, as they include or ignore to different extents issues of competence, tacit knowledge, etc.

(1993) that "finding out true things about the world is extremely easy [...] What scientists really care about are significant true things" (Weisberg and Muldoon, 2009, p. 229). They accept that agreeing on the source of significance is "an important and foundational debate in philosophy of science" (p. 229). Nonetheless, they state that their model remains agnostic about the source of significance judgement; they only require that "the community of scientists working on the same topic would make the same or nearly the same judgements about significance" (p. 229). It is better to be explicit here: from the structure of their model, it is clear that Weisberg and Muldoon mean that scientists agree on the *specific evaluation* of significance assigned to the truths that will be obtained if scientists pursue a specific approach, not just on general features of truths that contribute to significance or on some general method for assigning significance.[13] Notice that this introduces an assumption about the information scientists hold, of the kind that was criticised in the previous section; the criticism of assumptions in the model is presented in the next section.

The three components described above enable a definition of an epistemic landscape. The boundaries of the landscape are defined by the boundaries of the topic, such that each topic is described by a single landscape. The coordinates of the landscape are described by the different approaches available in the field; Each approach is a tiny "patch" of the landscape, with its neighbouring "patches" being the approaches most similar to it. The height of each coordinate (approach) is given by the scientific significance assigned to the truth that would be uncovered by following a specific approach within the specific topic (Weisberg and Muldoon, 2009, note 3 on p. 229); however, as mentioned below, and unlike the scientists in Kitcher and Strevens' models, individual agents *do not know* the significance assigned to each patch until it is visited, i.e. until some scientist actually pursues the approach. For simplicity purposes, Weisberg and Muldoon use a three-dimensional model, which makes the epistemic landscape appear as a mountainous region, which could potentially have peaks, hills, valleys, and plateaus. Weisberg and Muldoon present a graphical representation of such a landscape in their paper (Weisberg and Muldoon, 2009, Figure 1 on p. 230). A more detailed discussion of the components of Weisberg and Muldoon's model is presented in §5.1, where the components are incorporated into a revised version of the model used for computer simulations.

Epistemic agents, or investigators, occupy this landscape. The aim of investigation, of acquiring the most significant knowledge, is represented in this model by climbing to the highest point(s) of this landscape, and staying there. Weisberg and Muldoon's analysis of the division of cognitive labour is translated, by the use of the epistemic landscape model, to an analysis of the algorithms used by agents (scientists) to climb hills in such a landscape. In their model, Weisberg and Muldoon (unrealistically) treat the coordinate system of the landscape as fixed and known (to each agent), and the topology of the landscape as fixed but unknown (to each agent, though the topology is known from some "god's-eye" perspective). Their model focuses on two aspects:

1. Experimentation: how agents decide which direction to take, given the local topology,

2. Social interaction: how agents decide which direction to take, given the coordinates of

[13] Whether this can be achieved without the scientists also agreeing on the *source* or *criteria* for significance is unclear, though it seems plausible to infer that any model that assumes uniform agreement on specific significance assignments would also implicitly assume the existence of some mechanism for this state of affairs to take place. Such a mechanism is not specified by Weisberg and Muldoon.

other agents.

In Weisberg and Muldoon's model, agents only know the heights of the patches they have previously visited, the patch they currently occupy, and patches directly adjacent to the one they occupy, if these adjacent patches have been previously visited by another agent. Since individual agents do not know the height of a patch before occupying it, Weisberg and Muldoon avoid the "success function assumption" inherent in the lost-dog models, which they criticise (Muldoon and Weisberg, 2011). Since scientists also have limited "vision" about the position of other scientists in the landscape (i.e., they don't know which approach every other scientist is pursuing), they also avoid the "distribution assumption" criticised in the same paper.

2.4.1 Criticism of Weisberg and Muldoon's model

As in the case of Peirce's, Kitcher's, and Strevens' models, presented in the previous section, I am less interested in the conclusions Weisberg and Muldoon draw from their analysis. Rather, I am interested in their theoretical framework, and the substantive assumptions inherent in this framework. For the context of this thesis, the epistemic landscape model provides a great improvement over the models presented by Kitcher and Strevens. However, this new model also contains some problematic assumptions:

Scientists know and agree on significance While Weisberg and Muldoon have relaxed the assumption that all scientists agree on the difficulty of each problem, which means scientists don't need to agree on each approach's success function, they still assume a single, "objective", significance assignment to the results associated with each approach. Regardless of the origin of significance, it is unlikely that all scientists in the community would come to accept exactly the same evaluation of significance. Marked differences in significance attribution within a community could undermine Weisberg and Muldoon's results in two ways: first, if e.g. mavericks have very different significance assignment from followers, the followers may not follow, and second, it is not clear whose significance attribution should be considered when we evaluate whether a particular scheme of division of cognitive labour is better or worse. For example, in cases of scientific controversy, e.g. during the debate between "phlogiston" and "oxygene" chemical theories (Chang, 2012), scientists on different sides of the controversy are likely to more often (though not always) assign higher significance to research "on their own side" than to research "on the opponents' side".

Smooth landscape with few peaks In Weisberg and Muldoon's model there are two peaks, and the division of cognitive labour is assessed by the ability of scientist-agents to climb and explore these peaks. The smoothness of the landscape means that in the model, as approaches become more and more similar to some locally-optimal approach (the peak of a hill), so would their results become gradually more and more significant. Weisberg and Muldoon claim that their simulated landscapes represents a "common situation" in the target domain, but they do not provide any theoretical tools for constructing the landscape in specific situations. Thus, we may end up representing by a smooth landscape with few hills a reality in which small changes in approach lead to drastic changes in eventual significance, i.e. a rugged landscape with numerous hills and valleys. Such misrepresentation could in principle significantly undermine claims of optimal exploration, as strategies that work well on smooth landscapes may not work well on rugged landscapes.

This criticism may either be a criticism of an assumption, i.e. all landscapes are smooth, or a criticism of scope, i.e. that the model does not provide tools to help ascertain the correct ruggedness of the landscape.[14] Either way, a user of the model runs the risk of ending up with very misleading results.

Investigation does not change significance attribution In Weisberg and Muldoon's model the height of each patch, which is the significance attributed to the products of following the approach it represents, remain fixed while the scientists go up and down the hills. This is necessary, for example, for followers to be able to follow mavericks up hills – they need to still be hills by the time the followers get there. This assumption was already present in the lost dog models, and it was already marked for criticism in relation to them, by Kitcher himself. For example, significance could change over time as a result of research in other fields (an unexpected breakthrough), or because of a change in the environment (a problem with time-limited relevance, such as an epidemic). Recall, this is a problem because if significance does change over time, but we optimise *ex ante* significance without considering this change, the *ex post* evaluation of optimality could potentially deviate from the *ex ante* evaluation, such that what we thought would be an optimal strategy would turn out not to be. Instead, if we would have accounted for change of significance *ex ante*, the *ex post* and *ex ante* optimality evaluations would have agreed, leading us to *ex post* optimality, which is what we really want. The issue of dynamic significance attribution is the concern of Chapter 4.

In addition to the above worries, Weisberg and Muldoon's model also has shortcomings that make it unsuitable for direct application to the modelling of science funding decisions:

One topic per landscape As mentioned above, Weisberg and Muldoon's model identifies a single epistemic landscape with a single topic. Funding bodies, however, often have to decide on allocation of resources, or division of labour, between several topics, or a topic broadly construed. For Weisberg and Muldoon's model to apply to the decisions of funding bodies, the scope of "topic" needs to be relaxed, which may aggravate the problems of the assumptions mentioned above.

Unclear link between significance and well-being As will be discussed in greater detail in the next chapter, funding bodies aim to contribute to well-being. The connection between scientific significance, as it appears in Weisberg and Muldoon's model, and contribution to well-being, is far from clear. This means the model cannot be used to directly assess different funding strategies, because there is no way to compare the performance of the different strategies with respect to the key measure of interest, namely the contribution to well-being.

Due to the shortcomings of Weisberg and Muldoon's model, it cannot be used directly to critique current funding mechanisms and design new ones. The next chapter uses a causal picture of funding bodies' contributions to well-being to present a stepwise modelling strategy: in the first

[14] An example of a set of models that do include in their scope the tools for ascertaining the ruggedness of a landscape are the NK models of fitness landscapes discussed in Kauffman (1993, Ch. 2). In these models, a smooth landscape corresponds to low K, a rugged landscape to high K, where K is the number of different parameters that contribute to a particular increase in fitness of a particular organism. This issue is discussed in greater detail in the discussion of fitness landscapes in Chapter 3.

step, a novel model is constructed that assigns a well-being improvement measure to corpuses of public information. This model is then used in the second step, to create a revised version of the epistemic landscape model that assigns to each research project its potential contribution to the relevant corpus of information, if pursued. In this way individual projects are related to their eventual contribution to well-being, and the model can be used to evaluate funding strategies that select prospective projects based on their potential.

Chapter 3

Constructing new models of science funding decisions

We should fund science
that will make our lives better
eventually.

Introduction

The last chapter presented, and criticised, two generations of models of research-project choice. The first generation includes the works of Peirce (1879/1967), Kitcher (1990, 1993) and Strevens (2003). While promising, the economics-inspired framework shared by these models contains misleading assumptions about the nature of scientific research. The second generation includes the model presented by Weisberg and Muldoon (2009). While it successfully addresses some of the shortcomings of the first generation, it still falls short of providing a useful tool for thinking about science funding decisions, which is the aim of this thesis.

The key shortcomings of Weisberg and Muldoon's model, given the task at hand, were seen to be both conceptual and practical. At the conceptual level, Weisberg and Muldoon's use of the notion of significance, which plays a key role in the model, is largely left unanalysed, so that, for example, the various difficulties in assigning significance to projects is largely ignored. Furthermore, the delineation of the boundaries of the epistemic landscape is given very quickly, without pausing to consider difficulties in this delineation. At the practical level, there is very little guidance for the model user in how to gather information about the various model elements, including the mapping of distance between approaches, the delineation of the topic, and the assignment of significance to various approaches. A further practical limitation is the static nature of all the model elements, which assumedly was included for simplicity, but which prevents the accurate representation in the model of important features of the model's target domain.

Finding the existing models of project choice to be unsuitable for the design and critique of science funding mechanisms, this chapter develops new modelling tools to address this task. While informed by the models surveyed in the previous chapter, the model construction carried out in this chapter starts from a clean slate, in order to avoid the misleading assumptions contained in the existing models. One main source of problematic assumptions was identified in not having a clear enough concept of the benefits that arise from a "good" choice of projects.

The chapter therefore starts, in §3.1, with a clarification of the aim of public science funding bodies, and a high-level sketch of the causal pathways that lead from the funding decisions to the realisation of this aim.

The sketch of the causal pathway from the actions of funding bodies, i.e. project selection and allocation of funds, to the eventual benefits arising from these actions, is used in this chapter to guide a stepwise modelling strategy. The first model constructed focuses on corpuses of information, which are causally close to the aim of funding bodies, and then a second model is constructed, which focuses on project choice and relates funding decisions to information corpuses. A sketch of the first model, and clarification of concepts relating to corpuses of information, is presented in §3.2. §3.3 then presents a class of analogous models from population biology, the "fitness landscape", which provides useful machinery for the model construction.

The first step in the stepwise modelling strategy concludes in §3.4, providing a detailed sketch of a model of funding decisions and their effect on corpuses of information. This model, the "information landscape", allows a sharper definition of the question addressed by this thesis, and in principle provides the necessary machinery to evaluate alternative funding mechanisms. However, given the background from biology, it is deemed that the information landscape model is too complex to be effectively used for evaluating funding mechanisms. To show that this complexity is a feature of the target domain, rather than an artefact of modelling, the second step of the stepwise strategy presents a revised version of Weisberg and Muldoon's epistemic landscape in §3.5. This simplified model of funding decisions still carries, via its link to corpuses of information, important richness about the causal and interpretative complexities that influence the evaluation of funding decisions, and therefore offers a significant improvement over Weisberg and Muldoon's model. This revised epistemic landscape, unlike the information landscape, makes the complexity of the domain tractable and presentable in the model, as shown in the next two chapters where the model is used to represent the time-varying nature of scientific merit and to compare alternative funding mechanisms using a computer simulation. This tractable complexity of the simplified model then suggests the apparent complexity of the information landscape, which is even greater, is a feature of the domain of research project choice, rather than an artefact of modelling. If the domain of project choice is indeed complex in the way suggested by the models, such complexity will have important implications for the design of effective science funding mechanisms, as discussed in the last chapter.

3.1 Background for a model of science funding

When constructing a model that will help us critique and design science funding mechanisms, it is important that the model is capable of representing both desirable and undesirable outcomes we would expect from such a mechanism. We could then, using the model, compare different mechanisms, both actual and potential, and find out which mechanisms are most likely to produce more desirable outcomes and less undesirable ones. Thus, this chapter starts with an exploration of the aim of funding bodies, and a high-level account of the causal processes by which this aim is meant to be achieved.

3.1.1 The aim of public science funding bodies

What is the aim that public science funding bodies attempt to pursue when making funding decisions? As discussed in the first chapter, the rationale for allocating public funds to science is the expected payoff in terms of improved health, increased security, economic growth and improved quality of life. In light of this, it seems clear that public science funding bodies aim, or at least ought to aim, for an increase in well-being in the polity via the products of scientific research: new information, new techniques, trained individuals.[1]

Of the various ways public science funding can influence well-being, this thesis will focus on the generation of new, reliable, and communicable information as the core function of the scientific enterprise. The focus on the generation of new information, via research and publication, sets aside two broad and important aspects of research activity that continually make significant contributions to well-being: development of new techniques and training of individuals. Skills and new techniques can be transferred directly to commercial or public sector activity, e.g. via spin-offs, without going through the usual scientific publication channels. While the study of these processes is important, both in general and specifically for the evaluation of science funding mechanisms, they are going to be left out of the scope of this thesis. The motivation for the focus on published research is mostly pragmatic, and mirrors the decision to focus on basic research stated in the first chapter. While basic research projects can and do sometimes result in new techniques and spin-offs, it is often applied research and technology development projects that lead to these effects, whereas most basic research projects make their contribution via the publication of their results. As mentioned in the first chapter, these categories are porous, but there is an important underlying difference between *knowing that* and *knowing how* (Fantl, 2014): while some research projects contribute to well-being via informing the public about the way the world is, other research projects contribute to well-being by providing the public with new ways of doing things. While the sceptical argument presented in the following chapters could be further developed to bear on issues of technique development, this extension is not trivial, as it requires a discussion of the multi-faceted pros and cons of carrying out technological development work in academia versus in private industry. For lack of space the thesis will only focus on research projects that contribute by improving the information the society has about the world, via publication in a publicly accessible corpus of information.

Therefore, in the scope of this thesis, the following is assumed:

The main aim of public science funding bodies is the increase of well-being via the scientific generation of new, reliable, communicable information.[2]

That the above stated aim is true, or at least true enough, of public funding bodies, is a working assumption of this thesis. I therefore do not engage with positions that reject or challenge this hypothesis. Nonetheless, it might be useful to note a couple of alternative positions:

Knowledge for its own sake One alternative position, crudely presented, is that the scientific generation of information is inherently good, regardless of the effect new information will

[1] In the context of this thesis I will set aside complications arising from considerations of who should be considered a part of the polity, e.g. whether payment of taxes that support public research is a necessary and/or sufficient condition for inclusion in the relevant social group whose well-being is meant to be increased.

[2] A more detailed analysis of the aims of funding bodies is presented in Avin (2010). This earlier work requires, and defends, a stronger claim about the aims of funding bodies. For the argument put forward in this thesis, the weaker form presented above is sufficient.

have, or not have, on people's lives.

Cynical bureaucracies Another, cynical, objection, is that the real aim of public science funding bodies, like most large institutions, is merely to survive from year to year, preferably with an increased budget.

These alternatives are addressed, at least to some extent, by Kitcher (2011, ch. 5). If we accept a broad enough definition of well-being, or a rich enough picture of the constraints placed on the operation of funding bodies, these alternatives may be captured under, or be seen as not clashing with, the umbrella term of improving well-being. Since this point may be taken too far to imply that anything may be considered as promoting well-being, the next section clarifies the position of this thesis regarding the notion of well-being.

3.1.2 Agnosticism regarding theory of well-being

There are numerous theories of well-being, each with its merits and its criticisms.[3] Without going into details, it is expected that different notions of well-being will result in different evaluations of specific science funding mechanisms. However, this is one aspect of the problem of science funding that lies outside of the scope of this thesis, as described in §2.1. That section outlined different loci of questions regarding science funding: political, sociological, institutional, and epistemological. This thesis adopts the position that questions about the appropriate account of well-being belong to the political locus. It is assumed that, when coming to evaluate and design funding mechanisms using the account presented by this thesis, the evaluator or designer has already formulated an appropriate account of well-being to be used in this context. For example, a notion of well-being that relies on a democratic aggregation of individual and group preferences may be reached using the mechanism of ideal deliberation described by Kitcher (2011). Alternatively, a group of social scientists may present very convincing evidence that a particular list of objective criteria should serve as our concept of well-being. This thesis remains agnostic about whether either of these is a good approach to reaching an account of well-being in the context of science funding. By focusing on the epistemological aspects of science funding, this thesis brackets out complications arising from differences between accounts of well-being. The thesis assumes that if an account of well-being enables the evaluation of the products of research then some funding mechanisms are unlikely to perform as well as others in obtaining valuable results, and argues that epistemological factors bear on this likelihood.

This agnosticism regarding the theory of well-being will become relevant in a couple of places later in this thesis: later in this chapter, in the discussion of value assignment to the results of research, and in the next chapter, in the discussion of how these value assignments may change over time. In those places a more formalised and contextualised account of the above position will be presented.

3.1.3 A causal account of funding bodies' contribution to well-being

In what way do funding bodies' activities contribute to the well-being of the polity? To some extent, this would depend on the notion of well-being adopted by the model user. However, under many reasonable notions of well-being, a range of beneficial outcomes rely on, or are

[3] For a recent review of the philosophy of well-being see Crisp (2013).

improved by, the availability of reliable information. The role of scientists in this context is to generate and check this information, and to make it available to interested parties, such as other scientists, policy makers, various industries and the public. This kind of picture of information flow, from research to public information to actors who are in a position to increase well-being, is depicted (very schematically) in Fig. 3.1. The role of funding bodies in this picture is to support the scientists, and a good mechanism for supporting scientists will be judged by its ability to generate, over time, the greatest increase in well-being via the information channels sketched in this picture.

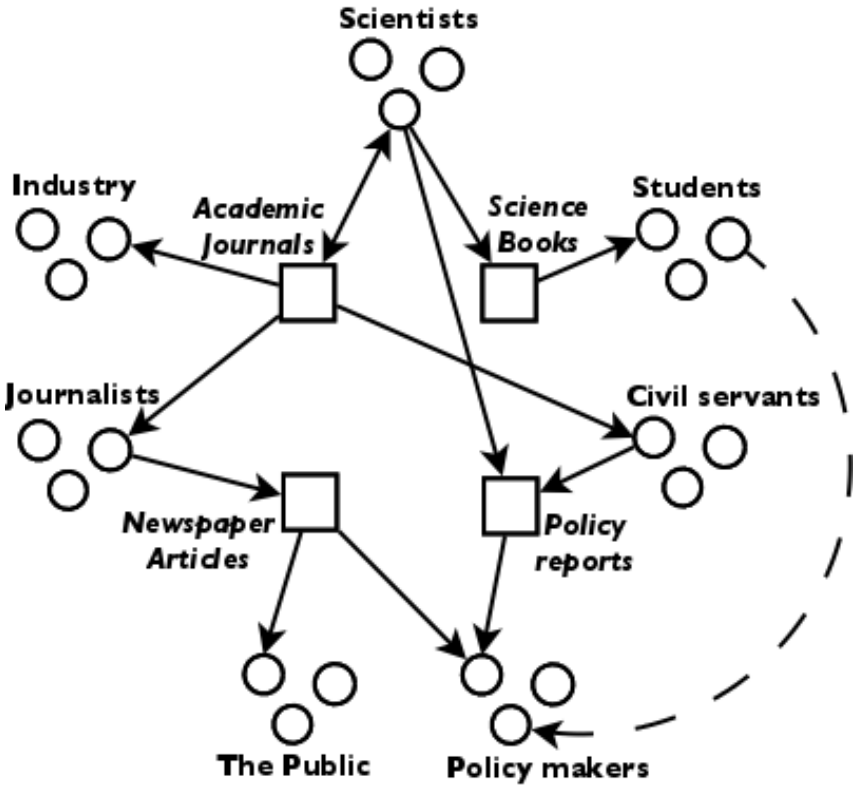


Figure 3.1: Graph representing a high-level abstraction of information flow in contemporary societies. Circles represent groups of humans, squares represent corpuses of information, full arrows represent the dominant directions of information flow. The dashed line between "students" and "policy makers" represents a relationship of "becoming"; similar relationships exist between "students" and other groups, but are left out for clarity.

To complement the above information-flow picture, it is also useful to consider the causal chain of activities involved. This thesis adopts a somewhat simplified picture of this causal chain, running from research proposals, through funding to research, and then through publications to activities that may increase well-being (see outline in Fig. 3.2). Which part of this causal chain should we focus our attention, and modelling efforts, on?

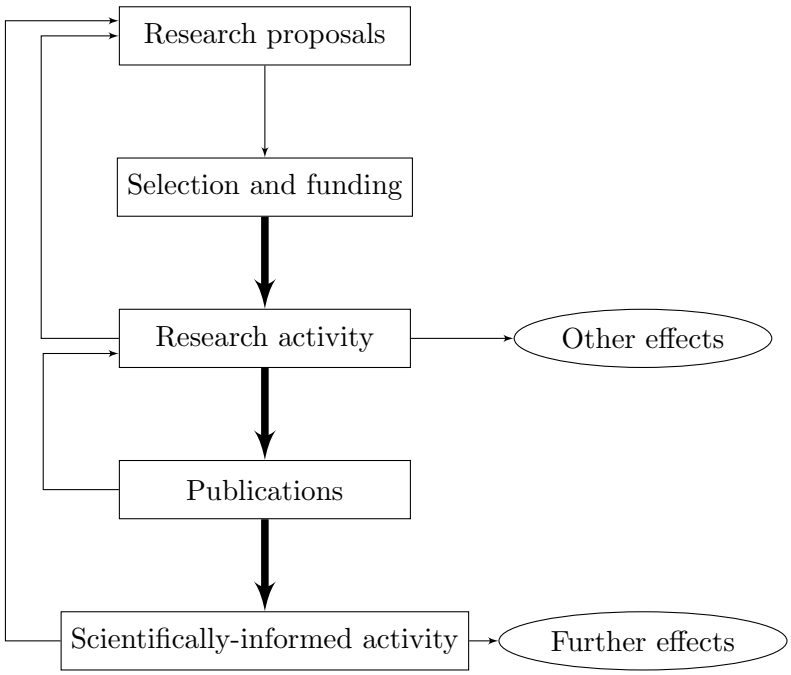


Figure 3.2: Schematic of the causal chain relevant to the design of a science funding mechanism. Within this (idealised) causal scheme, the inner causal chain from funding to scientifically-informed activity (bold arrows) comprises the core causal chain for investigation.

Given the picture above, there seem to be three candidate targets for modelling:

1. Scientific projects,

2. Corpuses of public information,

3. Activities that rely on scientific information.

The list progresses from the entities closest to the actual activity of funding bodies, i.e. the allocation of funds to research projects, to the entities closest to the aim of funding bodies, i.e. the scientifically-backed or supported activities (e.g. new products, better policy) that contribute directly to well-being. This chapter adopts a stepwise approach to model-building guided by this causal picture, as described in the next section.

3.2 Modelling the effects of science funding on public corpuses of information

The previous section focused the scope of this thesis on the aim of public science funding bodies of contributing to the polity's well-being by supporting the generation of new, reliable, communicable information via scientific research. Three possible targets for modelling have been identified: the research projects that could be supported, the public corpuses of information into which results from research are incorporated and on which scientifically-informed action is based, and scientifically-informed action itself. Based on this causal picture, this chapter adopts a stepwise strategy of modelling, where a first model links funding actions to corpuses of information. Then a second model is constructed that links individual research projects to a corpus. This approach is adopted as a way of minimising erroneous idealisations regarding the

link between individual projects and well-being, a fault that seems apparent in the models that go directly from projects to their effect (via the concept of significance or utility), as seen in the previous chapter.

This section presents the concept of a corpus of information as it is used in this thesis, sets out how corpuses are linked to contributions to well-being, and presents a sketch of a model which represents corpuses of information. Later sections then provide the modelling machinery required to expand this sketch into a more detailed model template that represents the causal relations between funding decisions and the resulting contents of corpuses of information.

3.2.1 Corpuses of information

Consider the collection of all actual scientific information stored by a given society at a given time in all formats, including books, journals, policy documents, lecture notes, etc. This collection makes up the complete information corpus of that society.[4] We can think about this corpus in the following manner. Imagine a world like ours, about which nothing is known; you can imagine it as an Earth-like planet covered in fog, or in complete darkness. Now imagine you have abundant time, and you view every piece of information available in the corpus. Every time you view a new piece, another region of the imagined world takes form: the fog dissipates, the darkness lifts, and the details uncovered match those described by the piece of information you are currently examining. When you finish examining all items in the corpus, you are left with a highly-detailed imaginary world. This imaginary world includes not only material components of this world and their descriptions, but also plans, motivations, laws, patterns, possibilities, and other non-material elements. This highly-detailed imagined world is a representation of the information contained in the entire corpus.[5]

While the complete information corpus may be the most comprehensive corpus, it is clearly not the most tractable. It is very hard to follow the effect that a single funding decision will have on the entire corpus of information available to a society. It is therefore useful to think of more restricted information corpuses. The restriction can be to any subset of the information held in the complete corpus; for example, a corpus can be restricted to the contents of an advanced manuscript, or the contents of a particular website, or the contents of a certain library, or the contents of all articles published in a specific collection of scientific journals of some discipline.

These corpuses persist in the society over time, though their contents change. To represent this change, we can index corpuses by a time, such that I_t is the contents of information corpus I at time t. At any given time the corpus will include all scientific contributions that have been made to this corpus throughout its history (except contents which have been deleted), and so at any given time various actors in the society who have access to the corpus can access all past contributions and base their actions on the information contained in these contributions. On this time-evolving picture, the aim of funding bodies is to influence the public corpuses of

[4] Motivated by the focus on communicable information (see §3.1.1), this definition of the information corpus does not include tacit knowledge. While issues of tacit knowledge, and particularly the generation of novel tacit knowledge, will certainly be relevant to a complete account of science funding, the complexities of the topic dictate that it be left out of the scope of this thesis, along with issues of skill and technique development (topics which are related to, but do not overlap entirely with, tacit knowledge). Reference works on tacit knowledge have been mentioned in Chapter 1, footnote 29.

[5] I am setting aside the complications that arise when we entertain the possibility that the collection of information can, and probably does, include contradicting items, and/or items expressed as a set of possibilities, with associated probabilities. A possible way forward, which I will not explore in any further detail, is to allow varying degrees of confidence in each aspect of the world, which will likely lead to the familiar Bayesian picture.

information such that at any point t they contain the information contents that are most useful for supporting actions that will increase well-being. A more formal statement of this point is given below.

3.2.2 Concepts of information

As with the concept of well-being discussed earlier in the chapter, the specific concept of "information" used in the model will be determined by the model-user. Nonetheless, it is useful to consider a few possible concepts of information to illustrate possible understandings of the concept of an information corpus.[6]

There can be at least two ways of conceptualising what information is in the context of a public corpus. One route looks at the role information plays in influencing decisions about actions. By taking this route we can account for the information in a corpus without understanding what it "means" to the members of the society: information is associated, through our observations, with particular actions taken by individuals or groups. Then, when a link between an item of information is associated with a particular action, a disposition towards this action is assigned to agents who have access to this item of information. On this view, "information" can be anything that has the power to influence such decisions or dispositions, i.e. the power to alter the probability of actions (and therefore also includes "misinformation"). This of course includes the ability to introduce new potential actions, or restrict the range of potential actions. We may choose, if we so desire, to limit the concept of information to cases where we have some reason to believe the actions of the recipient of information were desired by the sender of information (e.g. we may judge whether the sender has consequently become better off), and thus recover some notion of intentionality, still without assigning agent-perspective meaning to the information.

A different route starts from noticing humans are *conscious*, and therefore have a representation of their society and their environment. In such a case information is anything that has the power to modify the representation held by any agent, and intentional information is anything that was deliberately designed by an agent to influence the representations held by other agents. This can be related to the previous notion of information when we consider that actions rely on agents' representations of their environment. This does not mean the two routes are the same, though, as the nature of the relation between representations and actions, often considered under the umbrella term "rationality", is far from clear.

The specific concepts of information used in practice may be much sharper, e.g. based on individuals' beliefs, or the propositional content stored in their brains, or their conditioned responses to a set of stimuli. As mentioned above, the model developed will be left open with respect to the precise notion of information used, which in the context of this thesis will depend on the model-user's notion of information, and particularly scientific information.

[6]This brief discussion on information sidesteps a large and growing literature on the nature of information, and on the misconceptions that can arise from using the wrong metaphor to think about information. For an overview of different concepts of information, and different approaches to the philosophy of information, see Adriaans (2013). A specific account of information which is quite close to the position taken in this thesis is given by Collins (2010), in his discussion of "strings". An important list of caveats that should be kept in mind when thinking about information in the human context is given by Manson and O'Neill (2007).

3.2.3 Information corpuses and well-being

The relation between public corpuses of information and well-being is given by "epistemic fitness", a newly introduced concept which measures the extent to which the information available to a society fits its goals and values; some ways of conceptualising this "fitting" are discussed below. To be more precise, the "fit" is considered between the societal values and goals, as represented in the model user's adopted notion of well-being, and the *actual causal consequences of the existence of the information*. If we consider the totality of societal values and goals V to be represented by a particular desired state S (for example, a state in which all diseases have cures, food is abundant, etc.), a corpus of information I_A will be "fitter" than a corpus of information I_B if possession of the corpus I_A will (causally) lead the society closer to its desired state S than if it were in possession of the corpus I_B.[7] Since the full causal chain of consequences, resulting from the inclusion of a certain piece of information in the corpus, may take a long time to play out, we would often be considering a notion of "estimated fitness", which is some group's (e.g. funding bodies) best estimation of "actual fitness". The move from actual fitness to estimated fitness raises various concerns, and will be explored in detail in the next chapter. One aspect of that move that should be mentioned here is that assignments of estimated fitness to corpuses can be associated with a confidence level, i.e. a measure that describes how well-founded the estimates of fitness are taken to be for each corpus.

Given the novel concept of epistemic fitness it is possible to go back to the model presented by Weisberg and Muldoon (2009) and relate their notion of epistemic significance to epistemic fitness: a corpus of information will be attributed the highest level of epistemic fitness just in case it includes all and only information at the highest level of significance. As fitness relates to significance, it too is susceptible to the question of the origin of significance mentioned in the previous chapter. Given my position on the role of funding bodies stated earlier, the notion of epistemic fitness can be made a little sharper: information is "fitter" if it better serves the ability of actors within the society to promote well-being. While this is still a vague notion, it will be sufficient to produce a useful model of the decisions of funding bodies and their causal effects on well-being.

Unlike biological fitness (discussed below), which has a relatively clear definition (the ability to survive and produce offspring in a given environment), the conceptualisation of epistemic fitness will depend on the model user's conception of the source of values and goals among the human population. As far as this thesis is concerned the model produced will be left open with regards to the theory of well-being, as mentioned above. To state the point more formally, fitness assignment $F(I, W)$ is a function that takes two (many-dimensional) parameters, the corpus of information I as discussed above, and a specific, contextual measure of well-being W, and assigns to them a scalar f which is a measure of the "fit" between them: $F(I, W) = f, f \in \mathbf{R}$. In my analysis I will avoid, as much as possible, filling in a specific measure of W, and would therefore be discussing a family of models rather than one specific model. Of course, any specific use of the model will require the user to fill in W with an appropriate measure of well-being of their choosing. In the following I briefly present two perspectives on how to choose W, though I endorse neither – they are simply meant to illustrate how users of the model, with different concepts of well-being, may use the model differently:

[7] As part of bracketing out the political realm, this thesis sets asides complications arising from the fact that there are multiple ways to realise a desired state.

Values and goals are set by the environment From this perspective, the (non-human) environment is the main source of motivation for action, the origin of both dangers (in the form of predators, diseases, natural disasters, etc.) and well-being (food, raw materials, shelter, natural beauty, etc.). In this case, the most significant information is information about the environment, and specifically parts of the environment that pose threat or promise to the survival and well-being of humans. The most epistemically fit information corpus is the one which represents accurately all environmental threats and promises.

Values and goals are fully subjective From this perspective, what matters is the evaluation of "information products" by their consumers: academics, policy makers, industrialists, service providers and lay citizens all evaluate the information accessible and relevant to them according to their subjective preferences. Thus, the most epistemically fit information is the information that would best suit the (possibly subjective, possibly conflicting) values and goals of the consumers of "information products".

Just from the outline of these two perspectives, and the difference between them, we can see how the issue of epistemic fitness, and especially consensus agreement on epistemic fitness, is likely to be very problematic. In the case of the former position, the claim for an "objective" view of social goods may clash with democratic values, and brings into question the authority of whoever is making the assessment of these objective values (in our case, funding bodies) to make such judgements on behalf of the taxpaying public. In the case of the latter view, we run into the problem of social choice, or how to direct a joint activity that aims to satisfy numerous independent agents with varying preferences, which nonetheless operate in the same environment that, at least partially, contribute to their individual preferences. I do not presume to make any contribution to the debate about social choice or objective values in this thesis.[8] Rather, I choose to bracket out these worries into the social and political realm.

There is, however, an important aspect of the difficulty of reaching an agreement about the causal consequences of research projects and their contributions to well-being that cannot be bracketed out from the current epistemological consideration of science funding, and that is the possibility of prolonged, unbridgeable disagreement amongst those who decide on science funding allocation. This is the phenomenon described as incommensurability in Kuhn's account of scientific paradigms, and the situation that underlies some of Gillies' arguments against grant peer-review (see Chapter 1). I tend to agree with Gillies that in cases of incommensurability peer review serves as a negative mechanism by which the potential for novelty is reduced, because in that situation peer review becomes less of an informed comparison and more of a popularity contest between different schools. I disagree with Gillies on the point that this, in itself, is sufficient to undermine the use of grant peer-review entirely, because of the many situations in which there is no debilitating incommensurability that will undermine funding decisions. Rather, I argue that even when we exclude cases of incommensurability, and assume there is some mechanism by which decision makers can agree on a way to assign value to the expected causal consequences of different research projects, there are other kinds of uncertainty that will undermine the effectiveness of their decisions. Thus, from here on this thesis will only consider cases where there is general agreement about the state of the field of research (the corpus of

[8] For seminal works on social choice see Arrow (1963); Sen (1970). For a recent attempt to reconcile democratic values with project choice in science see Kitcher (2011).

information I) and the way to assign value to the products of research (the fitness assignment F).

3.2.4 Potential and counterfactual corpuses of information

When considering the decisions made by funding bodies and their causal effects on well-being, we need to consider various alternatives for each decision juncture. When considering past alternatives we would like to ask what contents our present corpuses of information would have held had the decision been different, i.e. we would want to consider counterfactual corpuses. Alternatively, for future corpuses we would like to know what information contents will be included in future corpuses if different choices are made, i.e. we would want to consider potential corpuses. Once we associate different decisions to different (actual, counterfactual, or potential) information corpuses, we can compare the epistemic fitness of each of these corpuses and use the relative epistemic fitness assignments to judge which of the corpuses is better, i.e. which makes the greatest contribution to the increase of well-being.

A model of information corpuses would therefore need to have the capability to represent, side-by-side, actual and counterfactual corpuses, or various potential corpuses. It will prove useful, in such a model, to represent the similarity between each pair of corpuses as a measure of distance between them. This similarity would be measured between the worlds the information corpuses represent; the more similar two imaginary worlds are, the closer they are. One way to achieve the distance metric is to use an intuitive notion of conceptual similarity between the worlds described by different corpuses of information, though the clear subjectivity of this intuitive judgement may jar with some readers. Another potential metric could be defined as the Levenshtein distance,[9] or some other edit distance, between the shortest description of the information contained in each collection, where the description is limited to some pre-specified formal language. I believe the intuitive conceptual metric is more natural and has fewer problems than the linguistic metric, but not much rests on this.

An alternative measure of distance, or rather of "nearness", could be given by considering the ease or difficulty of a transition from one corpus to another. Since the process of changing the content of a corpus is composed of many sub-processes, such as individual scientists entertaining the alternative, convincing others of the alternative, and the spread of the alternative view via various media, this measure of "nearness" will need to take many parameters into account. The technicalities of these points are discussed later in the chapter.

3.2.5 Sketch of a corpus-focused model

As sketched out above, a model of science funding decisions that focuses on corpuses of information should include representations of the following components:

- Actual and counterfactual or potential corpuses,
- A distance measure between these corpuses,
- An assignment of epistemic fitness to each corpus,

[9] Roughly, the Levenshtein distance between two strings is equal to the number of single-character edits required to change one string into the other.

- Time development, i.e. the progress over time from one corpus contents to the next given funding decisions.

Given such a model, it would be possible to compare the time-trajectories of different funding strategies, and compare, both at any given time and aggregated over a period of time, which strategies lead to corpuses of greater epistemic fitness.

Such model components and modelling method is analogous to a class of models in population biology called "fitness landscapes". Rather than re-inventing the modelling machinery required, the next section presents the "fitness landscape" model template, and various caveats about its use that have been gained over decades of experience and theoretical elaboration of the model parts.

3.3 The fitness landscape

Evolutionary biology looks at changes in the composition of populations. Two important processes that affect the composition of populations are reproduction and selection: organisms produce offspring which carry the same, or similar, genetic material; in parallel, organisms die due to natural selection, with a differential death-rate, so that some die faster than others. This gives rise to the concept of "fitness" in evolutionary biology, which (greatly simplifying and with much injustice to a complex field of research) I identify with the combined ability to survive natural selection and produce healthy offspring.[10] To visualise the effects of selection for large populations over long timescales, Wright (1932) has introduced the fitness landscape.

The exact status of the fitness landscape has been debated: some claim it is a model, others a metaphor or analogy, or even a simplified representation of reality (Plutynski, 2008). I follow Calcott (2008) in treating the fitness landscape as a model, or set of models, which falls under the framework of understanding models and model-related activity discussed in the previous chapter.

3.3.1 Structure of the fitness landscape

The fitness landscape visualises the fitness of all organism types, actual and potential, within a certain area of interest. Organism types can be characterised either by their genotype, the genetic material they carry and transmit through reproduction, or their phenotype, the manifest characteristics which determine their ability to survive and produce offspring. The different organism types are arranged side-by-side, either by genotype or phenotype, so that types which are more similar to each other are closer together. For example, if arranged by genotype, genome G will be adjacent to all genomes that are one-allele change from being identical to G.

Since a genome can have as many nearest-neighbours as there are genes in the genome, and all need to be equidistant, the resulting space is high-dimensional, with the number of dimensions determined by the number of nearest-neighbours of each genome. In the case of a fitness landscape of phenotypes, the number of dimensions is determined by the number of different characteristics that are used to evaluate a phenotype. The dimensions of the space can be discrete, as in the case of genotypes, or continuous, if the characteristics measured are continuous, e.g. height. This

[10] There are several different concepts, and measures, of biological fitness, along different distinctions: expected/actual; number of offspring/number of grand offspring; of tokens/ of types; etc. These various distinctions are discussed by Sober (2001). These differences are not relevant to this exposition.

high dimensionality creates interesting effects which were not noticed when the fitness landscape was originally introduced, as discussed below. In many visualisations of the fitness landscape, only two dimensions of type variability are represented.

To each point in the space, i.e. to each organism type depicted, a fitness value is assigned.[11] In the simplified case of two-dimensions of variability, the resulting model is a three dimensional landscape, with valleys representing a cluster of similar types that have low fitness, and peaks which culminate in a small cluster of types which, compared to their neighbours, are the fittest. An important feature of the resulting structure is the existence of local maxima: peaks which are higher than the neighbouring valleys, but lower than other peaks. Only the tallest peak in the landscape, the most fit type, is a global maximum; all other peaks culminate in a local maximum.

More formally, the fitness landscape is a scalar field over a potentially high-dimensionality configuration space, where individual dimensions of the space can be either discrete or continuous. Each point in the space represents a single organism type, either actual or potential. The scalar attributed to each point is the fitness of the organism type represented by this point (Stadler, 2002).

The model-representation, or visualisation, of the simplified three dimensional fitness landscape often takes the form of a contour plot or an isometric perspective of the three dimensional landscape, as in Fig. 3.3. Higher dimensionality versions are hard to visualise, but they can be simulated on a computer.

3.3.2 Movement in the fitness landscape

The main interest of the fitness landscape lies less in its structure, but more in the dynamics it helps visualise. Because the landscape represents all organism types of interest, it can be "occupied" by populations of organisms. Thus, we add a further attribute to the points in the space, of whether they are "occupied". Since many realistic situations exhibit groups that have similar phenotypes or genotypes (herds, kin groups), the resulting fitness landscape is occupied by "patches" or "clouds" of populations. Using the dynamics of evolution, we can predict the likely trajectory of patches, or the evolutionary future, of real or imagined populations.

Since fitness is a measure of reproductive success, if we sprinkle populations randomly across the landscape and fast-forward a single reproductive cycle, those patches that occupy valleys will shrink (low fitness results in high death rate), while those on peaks will grow (high fitness results in reproductive success). More interesting is the dynamics of patches that sit on slopes: the lower portion of the patch (lower fitness organisms) will have fewer offspring relative to those in the higher portion of the patch (higher fitness organisms). Thus, the lower part of the patch will shrink relative to the higher part. If, in addition, the population explores neighbouring regions of the landscape via accessible mutations, then in the visualisation, it will appear that the patch is "climbing" the slope. If we fast-forward many generations, we will see all patches either succumbing to valleys before they can get out of them, or climbing up their local peak and then thriving at its top. The resulting "equilibrium" state is for all occupied patches to rest on top of peaks (though not all peaks will necessarily be occupied), with no transition from one

[11] The use of a single scalar to describe fitness has been criticised by Ariew and Lewontin (2004). While I accept their criticism, I contend that it is still acceptable to use a single scalar in a simplifying model, while remembering that it is standing in for a family of different, case-specific, parameters of varying complexity.

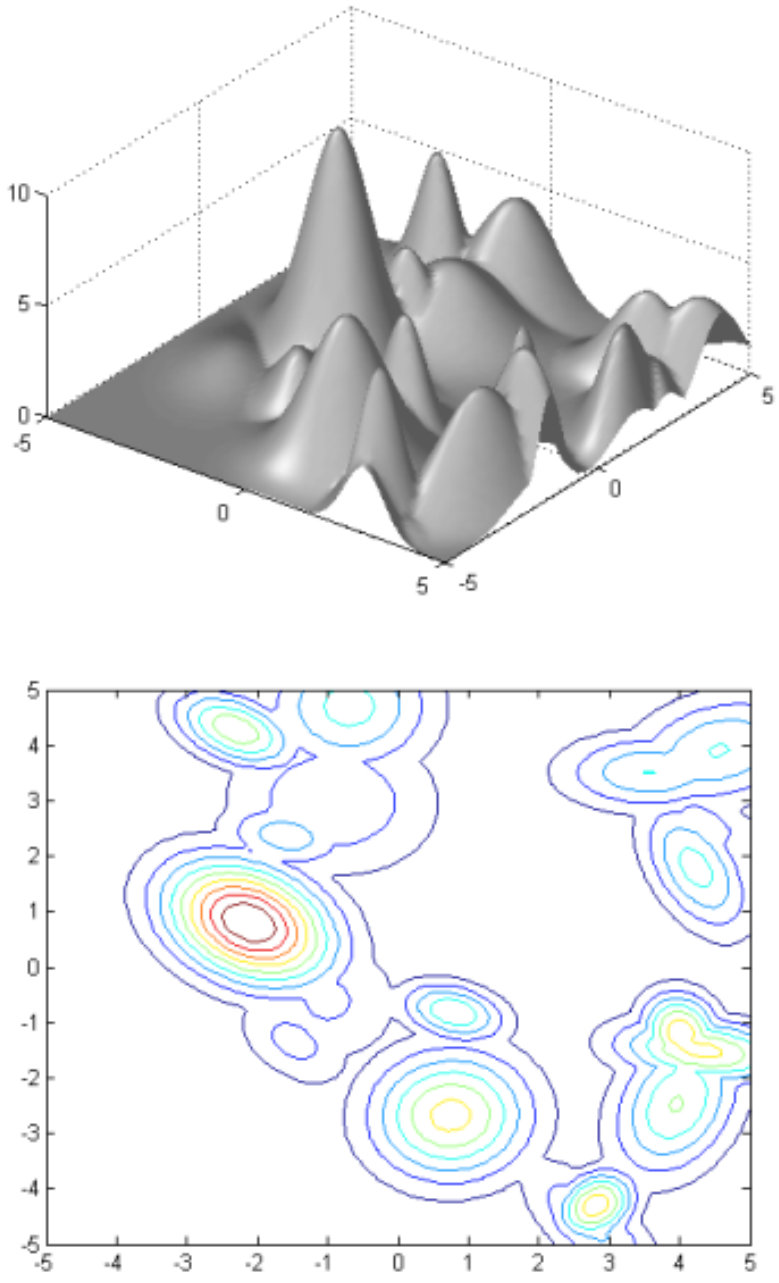


Figure 3.3: An example of common visual representations of fitness landscapes. The top image is a three dimensional representation of the landscape. The bottom image is the contour plot of the same landscape. The images are reproduced with the authors' permission from Yuan and Gallagher (2009). The method of their generation and use is described in Gallagher and Yuan (2006).

peak to another. This is why evolution is considered a "local maximisation algorithm", or a "hill climbing algorithm". This dynamic model is visualised in Fig. 3.4.

This simplified dynamic picture, which results from the three dimensional view of fitness landscapes and which leaves out various biological and ecological processes, has nonetheless proved very useful in the generation and formulation of theoretical questions in biology: How does this picture of patches climbing hills relate to the process of speciation? Can populations traverse valleys in order to go from a lower peak to a higher peak (the problem of "peak shift")? What happens to the dynamics if genes contribute to fitness independently or in large groups (Kauffman, 1993)?

It is the relative simplicity of the fitness landscape model, along with its ability to represent groups "exploring" a complex terrain, that make it appealing for use in modelling social phenomena in science. The similarities between this model and Weisberg and Muldoon's model, discussed in greater length later in the chapter, should already be apparent. However, experience with the fitness landscape model has also raised various criticisms and caveats about its use and limitations, which also prove relevant when the template is adopted for the context of this thesis. These criticisms and caveats are discussed below.

3.3.3 Caveats concerning the use of the fitness landscape

In recent literature the fitness landscape model has come under criticism for oversimplifying biological reality and misleading researchers (Gavrilets, 2004; Pigliucci and Kaplan, 2006). This criticism is concentrated around three focal points: the mapping from genotype to phenotype, the dimensionality of the landscape, and dynamics of the landscape itself. In addition, the notion of distance and nearness in fitness landscapes has been criticised by Stadler et al. (2001). While it may not be directly apparent, these criticisms are also relevant when similar models are used to model social phenomena, though some of the elements of the biological models (e.g. the phenotype/genotype distinction) take very different meaning, as discussed later in the chapter. These criticisms do not, however, lead to a rejection of the use of the fitness landscape; rather, they imply a certain list of caveats about correct and incorrect or misleading use of the model in various contexts.

Mapping from genotype to phenotype

In general, what contributes to the fitness of an organism type, in a particular environment, is its phenotype, its manifest characteristics. These can be directly observed, measured, and correlated with the number of offspring (assuming the population can be accessed and tracked over a period of time). For the observed specimen we can also experimentally obtain their genotypes, via sequencing or other methods. However, fitness landscapes often extend beyond the observed specimen, and represent *potential* phenotypes and genotypes. How do we extend the fitness landscape to include potential types?

First, we need to consider how to structure the configuration space for potential types. As stated above, this structure depends on the similarity between types. However, the "similarity" condition for equidistant adjacent points in a phenotypic fitness landscape is hard to specify: is a difference of one centimetre in height equivalent to a difference of one degree in beak angle? Conversely, equidistance is easy to measure in genomes, since all characteristics are, at least at some level of simplification, encoded in the same genetic medium.

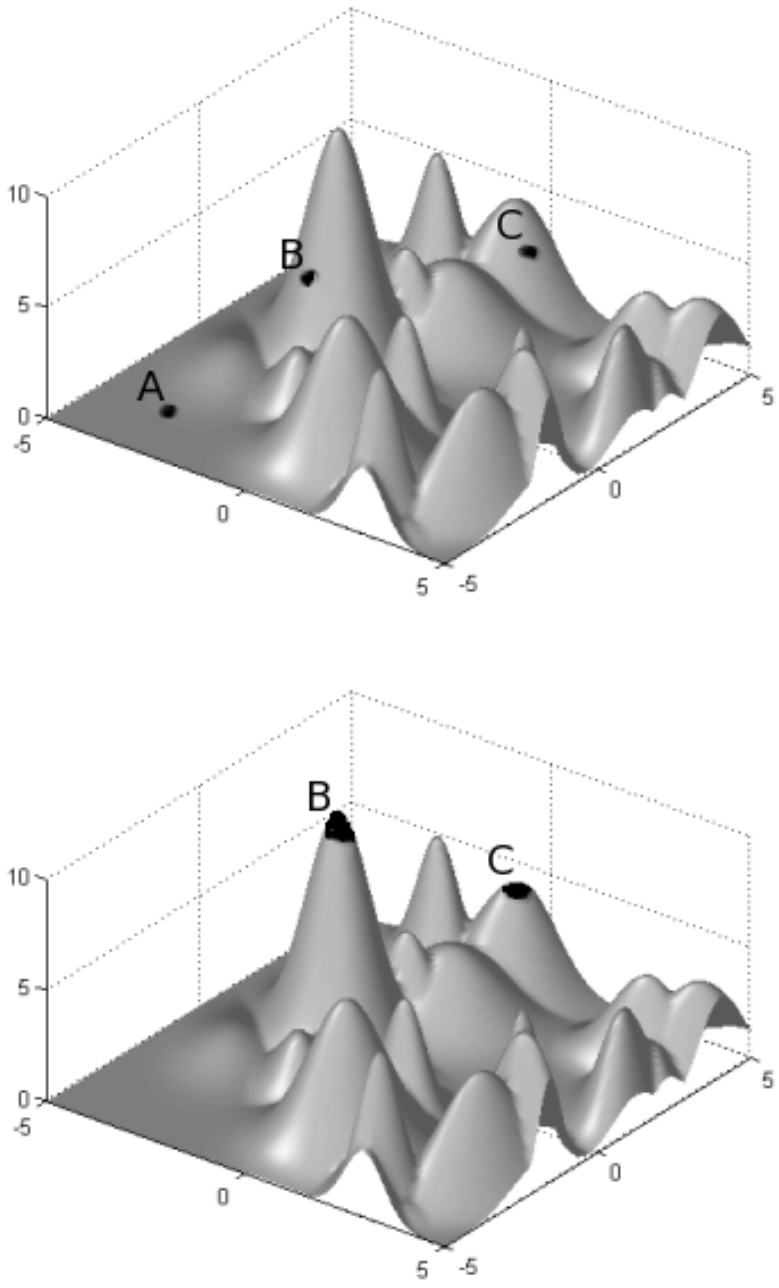


Figure 3.4: Visualisation of dynamics on a fitness landscape. Patches A, B, and C correspond to populations. The top image shows their initial positions in the landscape. The bottom image shows the result after many cycles of evolution: population A, which started at a low fitness location, has gone extinct. Populations B and C have climbed to their local maxima, which in the case of population B is also the global maximum, and have accordingly increased in size.

The major problem of extending the fitness landscape to potential types is the assignment of fitness to these types. Here, phenotypes fare much better: we can imagine the traits of a potential phenotype, and, based on our understanding of the physical interaction between the organism and its environment, predict (even roughly) how well it will perform. The case for genotypes is much more difficult, because it is hard, and sometimes impossible, to predict the phenotype that will result from a potential genotype. This derives from the problem of having no universal mapping from genotypes to phenotypes (Pigliucci, 2008). This criticism undermines the validity of hypotheses generated from genotypic fitness landscapes regarding the evolution of phenotypes.

Note that while this problem takes a very specific shape in the biological context, due to the encoding of phenotypes in discrete genomes, it raises a more general set of questions regarding the use of fitness landscapes and their various corollaries:

- To what extent can the model user assign fitness measures to the various coordinates of the configuration space, especially those which have not been directly observed?
- To what extent can the model user assign distances between coordinates, and do distances along different axes of the configuration space correspond to one another?
- To what extent are the two challenges above interacting in the target domain of the model, and if they do interact, by what mechanism do they interact?

The biological processes that influence the mapping from genotypes to phenotypes provide answers to the questions above in the case of biological fitness landscapes. It is reasonable to expect different answers will be given when the fitness landscape template is applied to other domains.

Dimensionality of the space

As mentioned above, the organism type space of fitness landscapes is known to be high-dimensional, sometimes with many thousands of dimensions, but is often depicted as two-dimensional (with fitness represented in a third dimension). Originally this was not considered to be a problem, because many of the interesting characteristics of fitness landscapes, such as valleys, peaks, and hill-climbing algorithms, are present in both low and high dimensions. The effect of dimensionality was considered in the restricted case of its effect on the ruggedness of the terrain (Kauffman, 1993). However, Gavrilets (2004) has shown, via computer simulation, that high-dimensionality landscapes are highly likely to contain a feature called a "neutral network". In the imagery of peaks and valleys, a neutral network is a network of ridges of similar height that connect distant peaks. In the presence of such a network, it is possible to travel from one peak to another without "going downhill" (reducing fitness). This feature of high-dimensional landscapes makes redundant one of the most interesting theoretical questions that arise from (low dimensionality) fitness landscapes, that of "peak shift", or how can populations travel from one peak to another without succumbing to selection on the way; in the presence of a neutral network, a population can travel from one peak to another via the network, without any need to suffer a temporary decrease in fitness.

While Gavrilets' observation has a dramatic effect on the problem of peak-shift, there are other interesting questions that arise from the fitness landscape which are unaffected by it (Calcott,

2008). Even the dissolution of the peak-shift problem does not make the fitness landscape a useless tool; in fact, we can now formulate new questions, e.g. do populations actually make use of neutral networks when travelling between peaks? Can this dynamic explain any of the observed phenomena in the biological world? It seems that the main upshot of Gavrilets' observation is not the rejection of the fitness landscape model, but a new appreciation that the dimensionality of the model is an important aspect of it, which may require some changes in how we represent it and work with it (Plutynski, 2008).

Dynamics of the fitness landscape

The ecological environment is not fixed: weather patterns change, other species arrive and start competing for the limited resources, prey goes extinct or migrates away, humans come in and start shooting everything they can. Because the fitness of an organism type is relative to its environment, a change in environment can cause a change in fitness. This means that for fitness landscapes to accurately describe their target domain, they need to be dynamic: not only do populations move across them, but they themselves change, with peaks turning to valleys and vice versa. The cause of the change can be external to what happens in the model (e.g. continental drift) or related to it (as prey evolves the fitness of the predator may decrease). Some effects might be bracketed out due to very long time scales, but since evolution itself happens on a long time scale, many environmental changes cannot be neglected. The influence of organisms on their environment in the context of evolution has been emphasised by a research program that looks at niche-construction, initiated by Lewontin and followed by various researchers (Odling-Smee et al., 1996).

Traditional fitness landscapes do not have a dynamic surface (Plutynski, 2008). Some have suggested the creation of fitness landscapes with dynamic, or "rubbery" terrain, but these are not in common use. So far, this remains as a known limitation of the model. Whenever environmental effects are considered important, other models need to be used.

Misleading notions of distance and nearness

In the fitness landscape model described above, the structure of the configuration space is the result of packing of the genotypes or phenotypes next to each other based on similarity. In the simplest case, genotypes are packed such that each genome is adjacent to all the genomes that are one single-point mutation away from it. In the case of phenotypes, these are often packed such that similar extensions are nearer than dissimilar extensions, so e.g. a beak length of 1cm is closer to a beak length of 1.5cm than it is to a beak length of 5cm, or a beak-less phenotype.

However, Stadler et al. (2001) have argued that this packing, particularly of phenotypes, is misleading. They distinguish two notions of distance or nearness, one which is simply a measure of apparent similarity, as described above, and the other a measure of the ease or difficulty of transition from one state to another. They argue that while the former is merely conventional, and can be quickly established to draw a fitness landscape for a particular case, it can easily be misunderstood for the latter, causing the users of the model to assume that movement across the landscape is equally likely in all directions. This assumption, they stress, is unwarranted given our understanding of the mechanism that underlies transitions in the phenotypic space, namely that they are the result of mutations of the genome.

In this highly technical paper, the authors show how accessibility of phenotypic states are:

1. of crucial importance to explanation of evolutionary phenomena, not easily explained by considerations of fitness and selection alone, and therefore not to be reduced to them mathematically, and

2. non-trivial in nature, e.g. they are not commutative (there may be an easy transition from A to B but a difficult transition from B to A), and when used to structure the fitness landscape, they result not in the familiar configuration space, but in a different and highly non-intuitive topological entity called a *pretopology*.

The paper describes the mathematical features of pretopologies that relate to evolutionary dynamics, and shows how they successfully explain various evolutionary phenomena. The authors focus on the phenotypic pretopology that arises from a genotypic configuration space. Thus, they take single point-mutations as being of equal probability in both directions and across the genome, giving rise to a simple transition rate that correlates to the Hamming distance between genomes, and the familiar configuration space when talking about genomes. They show how a phenotypic pretopology arises from the many-to-one relation between genotypes and phenotypes, namely from the existence of many point mutations that do not change the phenotype. The shapes and distribution of the networks formed by these "neutral" mutations determine the likelihood of transitions between phenotypes. The success of the pretopology, which relies on structuring "nearness" based on likelihood of transition, in explaining evolutionary phenomena, leads the authors to the conclusion that this should be the preferred notion of "nearness" in phenotypic fitness landscapes, and not the mere similarity notion of "nearness" that is often used. While discussing their advantages, it is useful to note an important disadvantage of pretopologies, not discussed by the authors: pretopologies cannot be visually represented, at least not with anything approaching the ease of representing "standard" configuration-space fitness landscapes. This tradeoff, between explanatory power and ease of representation, may explain why fitness landscapes are still mostly introduced using configuration-spaces and not pretopologies.

It is important to notice that the relevance of portraying the fitness landscape as a pretopology arises from the genotype/phenotype distinction, and from the fact that different forces operate at each level: at the genotypic level stochastic mutations (and the resulting configuration space) operate directly and influence dynamics, and the effects of selection are indirect; at the phenotypic level the effects of stochastic motion are indirect (giving rise to a pretopology), and to these effects the mechanisms of selection are applied directly, represented by the additional scalar "fitness" component. This condition will become relevant later, because the pretopology picture changes significantly as we substitute alternative interpretations for the phenotype/genotype distinction in non-biological systems.

3.3.4 Fitness landscapes and science funding

The discussion above of fitness landscapes, and caveats about their use, provides the necessary tools for constructing a model of scientific research, from the point of view of funding bodies. The key components mentioned in §3.2.5 can be identified with elements of the fitness landscape: different possible contents of a corpus of information are packed into a configuration space in the same way organism types are arranged, epistemic fitness is represented by the height of coordinates, a corollary of biological fitness, and the actual funding selection is represented by the occupation of coordinates, in the same way actual organisms occupy locations on the fitness

landscape. The details of the analogy, and the resulting model, are given in the next section, and they take into account the finer details of genotype/phenotype mapping, the dimensionality of the landscape, and the operative notion of similarity, all topics whose importance has been raised by considering the literature regarding fitness landscapes.

3.4 The information landscape

This section details the construction of the information landscape model, a model of science funding decisions that has corpuses of information as its main focus. This model will then be used in assisting the construction of a model of science funding that focuses on individual projects, while avoiding various misleading assumptions about the link between individual projects and eventual contribution to well-being.

I originally wanted to call this model "the belief landscape". However, the association between beliefs and assertions leaves out the vast, and increasing, amount of action-informing information that is transmitted using pictures, physical models, and, increasingly, videos. I wish to deliberately sidestep the question of whether informative videos *are* beliefs, generate beliefs, support beliefs, and which beliefs can/should be associated with a particular video, given which background information.

3.4.1 Structure of the information landscape

By analogy with the (biological) fitness landscape, the structure of the landscape is given by three components: the coordinates, the distance measure between the coordinates, and the scalar field that applies to these coordinates. A further element, not always discussed in the context of biological fitness landscapes but of importance to the science funding case, is the delineation of boundaries for the landscape in various use cases.

Coordinates

As described in §3.2, the corpus-focused model should contain a representation of different contents of corpuses of information, both actual and potential or counterfactual. For any given time t the information landscape represents a snapshot of all the possible contents of the corpus at that time, and so each coordinate of the landscape represents a *possible* alternative for the contents of the collection of public information under investigation (which may be the complete public information corpus or some subset thereof). Different dimensions of the landscape represent different ways in which different corpus contents can be "similar" or "near" each other, as described below.

Distance measure

§3.2 suggested that a model which focuses on information corpuses should include a distance measure between each pair of possible corpus contents. In the context of the fitness landscape template this functions as the metric for the landscape. I mentioned two possible ways of constructing such a metric. The first possible distance measure is one of similarity, using either intuitive conceptual judgements of similarity between the worlds described by the corpuses, or a

linguistic-representation focused notion of similarity using the edit distance between different descriptions of contents translated to a common language.

An alternative possible distance metric is one based on "nearness", or the difficulty of transitioning from one groups of contents to another. This measure may well end up being similar to "nearness" in biological phenotypes, which is not commutative (A could be "near" B without B being "near" A) and would thus require representation by a pretopology (see §3.3.3). However, unlike the biological case, the structure of the pretopology will be very hard to probe, because there is no clear genotype/phenotype distinction in science (Sterelny, 1994), and such a distinction is instrumental in mapping the pretopology of phenotypes in biology. For simplicity I will favour similarity-based distance. Some consideration of the difficulty of transitions is given in the next chapter.

Topology

In the same way that the height of each coordinate in the fitness landscape is given by the biological fitness of the phenotype or genotype associated with that coordinate, in the information landscape the height of each coordinate is given by the epistemic fitness of the corpus contents associated with that coordinate. Since epistemic fitness is a measure of the corpus contents' causal contribution to well-being, the exact assignment of height will depend on the model user's notion of well-being, and will also likely involve a certain amount of uncertainty, which may be represented as a confidence score associated to each estimation of epistemic fitness.

Boundaries

The boundaries of the landscape are given by the possible contents that can be included in a particular corpus. In the case of the complete public corpus such possibilities may be infinite, but more restricted corpuses (e.g. only the contents of a single journal) will place clearer restrictions on what information "belongs" in that corpus, and these restrictions will translate into boundaries of the associated information landscape.

3.4.2 Relating multiple landscapes

The model landscape presented by Weisberg and Muldoon was confined to a single manuscript. This boundary appeared in their work without justification, or any comment about which landscapes (manuscripts) merit our attention. This is unsatisfactory for the evaluation of science funding decisions, which often concern numerous topics and disciplines of various sizes, given the contemporary research agenda and available scientific knowledge (even if funding is restricted to "medicine", or even to "cancer research", there are many sub-divisions of the information within these areas) . In order to get closer to an evaluation of science funding decisions, we need a way to detect landscapes of interest, to understand which boundaries are most relevant, and to relate multiple landscapes to each other so that they may be considered simultaneously.

What we lack is a structure that will allow the combination of many bounded landscapes into one large model, while retaining the relevant independence of individual bounded landscapes, and at the same time represent the relations between topics that are required for evaluating decisions that cross topics. The conceptual construction of the information landscape as a landscape of the potential information content of corpuses, and of epistemic fitness as relating

to the causal consequences of information availability, provide useful tools for conceptualising varying boundaries and disciplinary links. As discussed above, the information landscape can be restricted to a corpus of any size, from a single manuscript to the entirety of recorded human knowledge, and so varying the bounds according to areas of interest is trivially feasible within the model, while retaining a rather sharp guideline for the boundary definition by relating it to the content of some actual corpus, rather than to a vaguer notion of a topic or a discipline.

The relative importance of corpuses, and the relations between them, are related to the causal chains that follow from the inclusion of information in these corpuses. An example of a highly-idealised schematic of causal information-flow was depicted in Fig. 3.1. Such schematics of corpuses, groups, and causal links enable us to explore the interaction between different corpuses, and therefore between different epistemic landscapes. Since these schematics portray corpuses that are available to the *same society* at the *same time*, I label the relationship between them "horizontal". If we wanted, we could just treat all the individual items in the different corpuses as members of one super-corpus, the "social corpus" or "unbounded corpus"; however, in doing so we would lose the specific information of the position of the corpus in the graph, i.e. the knowledge about the identity of the people relying on the information in each corpus. We would be better off adding items to the "social corpus" while remembering the specific intended users of each item, and weighing the epistemic fitness of the item according to the identity of its users. Thus, the fitness of information found in policy reports will be determined mostly by their usefulness to policy makers, whereas the fitness of the information in scientific journals (even in the same area) would be measured by their usefulness to other scientists, journalists, industry stakeholders and civil servants (and the many other groups not represented in this high-level abstraction). This means that the same item of information (say, a sentence or a numerical fact) can have different fitness, given the corpus within which it is located, and the identification of different corpuses and their causal influences facilitates the allocation of fitness to information items.

In addition to the horizontal relations discussed above, we may consider vertical relations, between corpuses that are part of larger corpuses. For example, the very high-level abstraction discussed above contains corpuses that in reality are a combination of many different smaller corpuses: "scientific journals" are represented as a single corpus in the graph, but in fact refer to many thousands of different journals, each one covering tens or hundreds of topics in hundreds of volumes and thousands of articles. For example, we can imagine a more detailed, but more restricted, graph, that only depicts information relating to the threat of rising sea levels resulting from anthropogenic climate change. Such a graph will still involve members from each of the groups of people depicted in the high-level abstraction, but the groups will now be much smaller and the similarities within groups will be greater. This means it will be easier to assign fitness to particular items in corpuses of this more specific graph. We can then aggregate the fitnesses of particular items in several smaller corpuses (e.g. articles on the threat of rising sea levels, articles on desertification, articles on extreme weather events, etc.) to gain a better understanding of the fitness of a higher-level corpus (all articles on climate change). Again, the vertical connection enriches our understanding of the fitness of individual items of information in the higher-level corpus by retaining information about use and users from a more detailed picture.

3.4.3 Using the model for evaluating science funding decisions

As stated at the beginning of this chapter, funding bodies aim to increase well-being via scientifically generated knowledge. The information landscape model associates a bounded landscape of epistemic fitness to each information corpus available to the society, where epistemic fitness in turn is causally related to well-being as described above. Thus, funding bodies aim to increase the overall fitness of the information contained in the society's corpuses. If we have a global notion of well-being, either via objective values or an aggregate of subjective ones, we can recombine the bounded information landscapes to a single, universal information landscape, each point of which represents the totality of information available to the society, and the fitness of each coordinate representing the fit between the totality of information and the various information needs of the society. Combining all of the corpuses which are under the auspices of a particular funding body, we arrive at that funding body's relevant information landscape.

Given the theoretical framework developed above, the central question of this thesis, i.e. *Are the processes used by public science funding bodies to make funding decisions rational, and can they be made more rational?*, can be restated as:

Are the processes used by public science funding bodies to make funding decisions likely to lead us to coordinates of the relevant information landscapes with higher epistemic fitness?

This question is tackled in the remainder of this thesis.

3.4.4 Complexities of the information landscape

While the sketch of the information landscape presented in §3.2.5 is relatively straightforward, the detailed investigation of the model components and the background given by biological fitness landscapes show the model to be, at least potentially, very complex.[12] This section presents a summary of the sources of complexity in the model, and some implications for the study and design of science funding mechanisms.

High dimensionality

Given the mechanisms of representation borrowed from biological fitness landscapes, and when relying on intuitive conceptual judgements of similarity between corpus contents as the landscape's metric, it is clear that the information landscape for any corpus that is rich enough to be of interest will require a high-dimensional representation. Many dimensions are required to depict all the different ways that potential or counterfactual corpus contents can differ from other corpus contents, as corpus contents contain information about many objects, many features of those objects, many interactions between these objects, potentially many laws that govern the behaviour of these objects, etc. Each of these numerous aspects of the domain can be potentially explored and learnt independently (though of course many studies will include the study of more than one aspect of the domain) and therefore each gives rise to another dimension along which to differentiate potential corpus contents. As mentioned in §3.3.3, landscapes of high dimensions give rise to different dynamics and exploration strategies than do low-dimensionality landscapes,

[12] The term "complex" here is meant intuitively to indicate a system with many interacting components, leading to either limited predictability or to a large amount of effort required to make good predictions. For a review of the philosophical study of the concept of complexity see Immerman (2011); Zuchowski (2012).

and these dynamics and strategies are hard to elucidate. Simple intuition is usually insufficient and often misguiding in understating the time-dependant behaviour of such landscapes, requiring the use of computer models.

Wide range of disciplines

A construction of an information landscape that accurately captures interesting features about the domain would require the model user to draw on knowledge and skills from a wide range of disciplines. One aspect of this is trivial, when dealing with corpuses that span many target domains: a prediction of potential contents and estimation of fitness for these corpuses will require knowledge in all the domains covered by the contents of the corpus. Another aspect is the importance of psychological and social aspects, in addition to empirical knowledge about the target domains (Wilkins, 2008).[13] Psychological and social aspects can play a role in the model in influencing the metric: transitions between corpus contents are important for a time-dependant simulation of strategies that aim to reach higher fitness, and these transitions would depend on the psychological and social appeal of the contents to scientists and other audiences. In addition, psychological and sociological aspects are also likely to play a role in the accurate assignment of fitness to potential corpus contents. While these aspects have been explicitly "bracketed out" from the model itself, it is clear how they present an additional burden for the model user who wants to use the model to design a better funding mechanism.

Causally removed from direct funding actions

As mentioned above in the description of the stepwise strategy, the focus on corpuses is motivated by its causal proximity to the aim of funding bodies, i.e. the availability of information that will contribute to increases in well-being. However, the causal proximity to the desired effects comes at the cost of causal distance from the funding decisions themselves, i.e. the selection of individual projects and the actual research activity in the lab or the field. One way this causal distance is a source of complexity is in the non-trivial relation between research projects and research results: a single project may produce multiple results, and a single result may be the product of many projects. The information landscape focuses on corpus contents, and these contents contain results, not projects. In contrast, funding bodies select from proposed projects, not from proposed results. This kind of non-trivial relation is similar to the relation between genotypes and phenotypes in the biological case, suggesting for example that what might be considered a straight-forward configuration space is in fact a pretopology (see §3.3.3). Such structures take us further from our intuitions about the behaviour of the system and require computer simulations.

Implications of complexity

The above sources of complexity combine to suggest that the information landscape may end up being quite hard for model users to use. This difficulty could have two, quite different, interpretations. The first interpretation is that the complexity is merely an artefact of the

[13] Wilkins (2008) presents a different application of the fitness landscape model template to scientific activity, though he is committed to Hull's view that takes selection of organisms in biology and selection of ideas in science to be the same process (Hull, 1988), a view not endorsed by this thesis. For a criticism of Hull's and Wilkins' view see Sterelny (1994).

model construction. Indeed, it is not difficult to construct very complex models of very simple phenomena. In such a case, the modelling effort presented so far has been merely a waste, a dead-end of modelling. However, the more interesting interpretation is that the target domain itself is complex, and that this complexity is significant for the decisions of science funding and their eventual contribution to well-being.

To argue that the latter interpretation is the correct one, this thesis moves from a model of corpuses to a model of individual research projects, similar to Weisberg and Muldoon's model presented in the previous chapter. Using this, more simplified, model of the target domain, the thesis will argue, over the next two chapters, that the target domain is indeed complex, and that this complexity can be captured in this more simplified model. This then strongly suggests that the complexity of the information landscape is indeed indicative of the complexity of the problem of choosing a strategy for science funding. This complexity then carries significance in our choice of funding mechanism, as described in the last chapter.

3.5 From the information landscape back to the epistemic landscape

The information landscape model presented above was constructed as a way of assigning fitness values to hypothetical contents of a corpus of information, and displaying them side by side with distance given by the level of similarity between the corpus contents. This model gives us a fairly comprehensive understanding of what we want from science: we want it to move us towards corpuses of higher fitness. This abstract aim can be translated into actions in various ways, depending on the fitness assignment method we adopt. Nonetheless, it is likely to include some weighted or contextualised subset of the following: the removal of pernicious falsehoods, the discovery of useful and interesting phenomena, the development of applicable and explanatory theories, and the attainment of practical knowledge about the implementation of theories to areas that matter to us.

The return to Weisberg and Muldoon's epistemic landscape, or a version thereof, described below aims to connect this general aim, of increasing the fitness of the information corpuses available to the society, with the specific actions of the research community. These actions are: selection of projects, allocation of funds, simultaneous activity of research on multiple projects, and attainment of results (with resulting changes in fitness) (see Fig. 3.2). These processes are very hard to track using the information landscape, as they relate to *scientists* and *projects*, not to *corpuses of knowledge*. This difficulty can be addressed by shifting the modelling focus from the information landscape to the epistemic landscape. The justification for this shift, and the bridge principles required for it, are described below.

3.5.1 Bridging the information landscape and the epistemic landscape

How is scientific research practised? At a very high level of abstraction, we can describe the practice as a social activity, where at any given time a group of selected individuals, the investigators, pursue various research projects. The aim of these projects is to make some contribution to the corpus of information available to the scientific community, to interested parties, and to society at large. When a research project ends, its results are incorporated into the relevant corpuses, which means the contents of these corpuses change. The process by which

results enter a corpus is of course complicated, and philosophically interesting in its own right, but falls outside the scope of this thesis. We can "black box" this process, and assume, for the sake of argument and with a high level of idealisation, that its function is to prevent changes to the corpus that lower its fitness.[14] Thus, we have a picture of various projects, each pursued by one or more investigator (or group of investigators), and when any of these projects end, a change in the corpus of information occurs, associated with a non-negative change in fitness.

We already know how a change in contents of a corpus is represented: it is a transition on the information landscape. The different available projects at a given time, or rather the different *results* the society might obtain at a given time, correspond to allowed transitions on the information landscape. The large number of possible results means that the dimensionality of the information landscape is very high. The connection between results and corpuses can also be conceptualised in the other direction: given a corpus of information, and its associated information landscape, the set of all one-step transitions which do not decrease the fitness of the corpus are acceptable candidate *results* for scientific projects, and each can be assigned a *fitness differential*, which is the (counterfactual) non-negative *change* in the fitness of the corpus that would occur were the result incorporated in the corpus.

In the biological genotypic fitness landscape, all the single-edit neighbours of a genotype are considered to be independent of each other, as they represent changes in different genes, and so changes are represented as occurring on directions that are orthogonal to each other. In the information landscape, however, changes can relate to various aspect of the target domain that are not necessarily independent of each other. For example, a corpus may contain descriptions of enzyme activity, and some of the enzymes may share structure and functional element. Thus, we can conceptualise changes in the information landscape as being non-orthogonal, with the angle between directions determined by the amount of overlap between characteristics in the target domain, with fully independent features of the domain being orthogonal, and overlapping features being assigned an acute angle. This conceptualisation also suggests a metric for the distance between results: two results which shift the corpus in a similar direction (i.e. the angle between the direction of the change is small) would be closer to each other than two results which shift the corpus in different directions (i.e. when the angle is large).

Given a metric for distance between results, and a scalar field that applies to all results, we have a definition for a landscape of results. This is the epistemic landscape. The epistemic landscape is the landscape of all results that, if added to the corpus of information, would bring about a non-negative change of fitness for that corpus (as mentioned above, negative contributions are assumed to be transformed into zero contributions by the checking mechanism). In the following chapters I talk mostly of projects, not results. I consider this to be shorthand:

[14]The abstraction here is twofold: first, the role of the mechanisms for the acceptance and distribution of scientific results is taken to be merely the prevention of errors, or otherwise unfit results, where in fact we know the roles of these institutions are varied, dynamic and complex; second, we choose to fill in the "values" component of the fitness evaluation in a way that is broadly positive of science and its historic successes (though not trivially so). This position is then farther abstracted for simplicity's sake, into taking these institutions to be infallible in this restricted role, when in fact we know this is not the case, and various false or otherwise unfit results have in the past been incorporated into the scientific corpus. Nonetheless, from a broadly pro-science perspective it seems reasonable to assume that overall the institutions for accepting and disseminating new scientific results *do* check these results for errors and that, to the best of society's abilities at the time, they often get it right. The motivation for adopting a pro-science position, for the sake of argument, is threefold: first, it is often the position of those who are in charge of the system of science funding as it is now (see Chapter 1); second, the sceptical arguments against the current best practices, which are developed from the detailed analysis in this chapter, carry most force against those who support the current best practices; lastly, it is the position of the author.

a project is the pursuit of a certain result. This involves a known simplification, as discussed above: any real mapping from projects to results is unlikely to be one-to-one, and the level of uncertainty regarding projects and results is likely to differ, sometimes significantly (we know what we're looking at but not what we'll find). Where possible, I address this simplification directly, e.g. with a visibility condition (see Chapter 5). Otherwise, the epistemic landscape model is taken to represent both the results, with their associated fitness differential, and the projects that lead to these results. A completion of a project on the epistemic landscape is therefore directly and uniquely associated with a transition on the information landscape, thus providing the bridge between the two models.

It is important to note that the bridge also operates in the other direction. When a project is finished (on the epistemic landscape), its result is incorporated into the corpus, which is associated with a transition in the information landscape. This transition brings the corpus to a new location, which has a different set of transitions it can undergo which are associated with a non-negative change of fitness, which means a different set of candidate results, and a different set of projects. Thus, the completion of a project on the epistemic landscape leads to position change in the information landscape, which leads to a change in the epistemic landscape itself (coordinates and associated scalar field). This complex interaction is what leads to the appearance of *fitness dynamics*, as discussed in the next chapter.

3.5.2 Advantages over existing models

The stepwise modelling strategy, via the information landscape model to a new version of the epistemic landscape, was chosen to avoid various erroneous simplifications or limitations that were observed in the models surveyed in the previous chapter, where modellers moved directly from beneficial outcomes (as significance or utility) to the projects themselves. Of the models presented in the previous chapter, the most promising was Weisberg and Muldoon's model of the epistemic landscape, though it too was deemed unfit, as it currently stands, to meet the requirements for evaluating and designing science funding mechanisms. Recall my criticism of their model, which was presented in the previous chapter: In Weisberg and Muldoon's model,

WM1. There is no clear connection between approaches and well-being: significance, which is the measure Weisberg and Muldoon assign to each approach, is an external parameter, with unspecified origin and unspecified operationalisation;

WM2. Significance is single-valued, there is no difference in individual scientists' evaluation of significance;

WM3. The act of investigation, represented by a scientist occupying a patch, does not change the significance attribution of any approach;

WM4. There is a one-to-one assignment of landscapes to "topics", without any representation of links between topics/landscapes or ways of combining them;

WM5. The epistemic landscape is smooth, having few peaks and valleys.

The main advantage of my revised version of the epistemic landscape is in addressing WM1, by making a clear causal link between research projects and well-being, by requiring a model user to specify a method for assigning fitness to information corpuses. This assignment of fitness then

translates into an assignment of a *fitness differential* for every result in the epistemic landscape. This link also influences WM2 and WM3: by linking fitness to real, albeit very complicated, actions, we can start asking questions about how scientists get to know about these actions, and how their own actions influence the causal chains that determine the fitness of a particular corpus. I addressed WM4 by introducing the notion of a hierarchy of landscapes. WM5 is not addressed directly by the current model, but is explored in the simulations presented in Chapter 5.

Using the wider-scope and higher-fidelity models developed in this chapter, the next chapters show how they can be used to explore complex processes that occur during research activity, and the influence these processes have on the expected effectiveness of various funding mechanisms.

Chapter 4

Dynamics of epistemic fitness

Each project's impact
will be determined by a
complex causal mess.

Introduction

The previous chapter presented a revised version of the "epistemic landscape" model as a tool for evaluating and designing science funding mechanisms. The model was created by reference to an intermediary model, the "information landscape", which represents a society's assignment of epistemic fitness to a range of possible corpuses of information. In these models, epistemic fitness is the causal contribution of scientifically generated information to the well-being of the polity, where the notion of well-being is left open to be specified by the model user.

This chapter and the next present one way in which these models can be used to evaluate the rationality of funding strategies. The aim of these chapters is to present a sceptical argument, challenging the ability of funding bodies to know *ex ante* which of a set of proposed projects will contribute most to well-being. This sort of sceptical argument can take one of at least two lines: one, to challenge the capacity of funding bodies to gather the relevant information at the time of making their decision; the other, to challenge the capacity of funding bodies to predict how the fitness of projects will change in the time between project selection and completion of the project. Since the argument is sceptical, it is enough to show that *either* of these challenges undermines the capacity of funding bodies to successfully pursue an effective funding strategy via prospective evaluation of projects; conversely, if one were to defend a particular strategy of this kind, one would need to meet both challenges. This thesis will only look at the latter of these two options, bracketing out the first challenge by assuming, unrealistically, that funding bodies have access to all the knowledge that is available at the time of decision making.

As mentioned in the previous chapter, public corpuses of scientific information persist in time, but their contents change over time. These contents change as the result of ongoing research activity, which in turn is supported by public science funding. Since multiple research projects can contribute to the same corpus, the increase in the fitness of the corpus resulting from a single project (the project's *fitness differential*) will depend on other projects that have finished in the past and contributed to the same corpus. This can be represented in the epistemic landscape model as fitness dynamics, i.e. as continual changes to the fitness differentials assigned to the

projects that make up the landscape as the community of investigators explores the landscape and, by this exploration, continually updates the associated corpus of information. The outline of the theoretical importance of fitness dynamics, its relevance to the rationality of strategy choice, and a discussion of the subtle differences between estimated fitness and actual/counterfactual fitness, are presented in §4.1.

This chapter demonstrates various processes by which fitness can change. §4.2 presents brief historical examples of episodes in the history of science where fitness changed dramatically over a relatively short period of time. The examples are chosen for the clarity with which they show a change of fitness, but this clarity comes at a cost of rarity: the situations described display drastic effects, and are therefore likely to be rare occasions in the history of science. A possible counter-argument, in defence of the ability of funding bodies to maximise contributions to well-being via prospective evaluation, would then be to argue that the processes presented in this chapter would rarely interfere with the strategy, and therefore in most cases the strategy would be successful. To meet this challenge, I turn to the models developed in the previous chapter. The theoretical sections of this chapter that follow the historical examples present, *qualitatively*, how different dynamic process of fitness change could be represented using the epistemic landscape model. This qualitative modelling is then used, in the next chapter, to make a quantitative computer simulation that shows the dynamics of fitness in a system where several processes co-occur, with varying probabilities and magnitudes.

The chapter concludes in §4.6 with a consideration of some of the wider influences of fitness dynamics, both specifically on existing works in the philosophy of science and on models of scientific activity in general.

4.1 Outline of a dynamic picture of epistemic fitness

Funding bodies use peer review for merit evaluation because they want to fund the most meritorious projects, under the belief that picking the most meritorious projects will advance their aim of increasing well-being via high quality research. This behaviour is consequentialist: the choice of a certain strategy (peer review) is based on the consequences of that strategy. A strategy is preferred because it will lead to more positive consequences, in this case a greater increase in well-being. Thus, the challenge that dynamic fitness presents to funding bodies can be summarised as:

How do you know, once you decide to employ a strategy that you predict will lead to beneficial consequences, that the actual consequences will still be beneficial when you attain them?

It is trivial to see why this challenge does not arise if the benefits associated with various alternatives are considered to be fixed.

4.1.1 Fitness dynamics and consequentialism

The challenge presented above is similar to a general worry for consequentialism, described by Burch-Brown (2012, 2014) as the "objection from cluelessness". It would be beneficial to consider some of the similarities with this objection, and its counters, but also some key differences. The objection from cluelessness, which Burch-Brown develops primarily from Lenman (2000), states that the unforeseeable consequences of our actions greatly outweigh, in number and scope, those

consequences which we can foresee. Therefore, the actual benefit of taking one course of action over another will largely depend on information not available to us, information we are clueless about. This means that overall, if actual consequences are what matters, and we seem to have a strong intuition that this is the case, then consequentialism cannot recommend any particular action over any other. Specifically, Lenman argues that choosing the act that will maximise foreseeable outcomes is equally justifiable as choosing at random, or choosing the act that will maximise foreseeable harm.

My own sceptical argument, which stems from the dynamic nature of epistemic fitness, also undermines the rationality, or justification, of certain strategies, due to the weight of unforeseeable outcomes. In the case of dynamic epistemic fitness, however, the scepticism is not global. For example, while I argue that choosing projects at random may (in some cases) fare as well as trying to pick the best projects, my analysis still supports the evaluation that both of these strategies will fare better than doing no research, and better than picking the worst projects. This is because the source of cluelessness is not vague and all-present, as in Lenman's argument, but stems from specific processes of dynamic epistemic fitness, as outlined in this chapter. Thus, Burch-Brown's counter to Lenman's sceptical argument, based on the justification we can give to *strategies* if not to specific *acts*, does not counter my own, because, as I show in this and the next chapter, the strategy of maximising current-best projects fails, in certain contexts, *given what we know about science*.

4.1.2 Estimated versus actual epistemic fitness

Epistemic fitness, as defined in the previous chapter, is a measure of the extent to which the *actual* causal consequences of including a particular information item in a particular information corpus will bring a society closer to its collective goals and values. This raises two concerns. The first worry is that the information landscape includes many *potential* items of information, and items that are not included in the corpus do not have causal consequences. We can make some headway by replacing *actual* with *counterfactual*, i.e. epistemic fitness measures the contribution of the causal consequences that would take place if the item of information was included in the corpus. This worry is also present in the biological fitness landscape, where we try to assign fitness to unobserved phenotypes (see §3.3). As in the biological case, our best response is to try and extrapolate to the best of our ability from the known to the unknown, and remember that the explanatory and predictive power of hypotheses that rely on parameter values within the unobserved domain is, at best, limited.

The second worry is that the totality of causal consequences, both actual and counterfactual, may take (or would take) a long time to play out, and until they do we will not know what they are going to be – indeed, this is the main point of this chapter. How, then, are we to assign values of fitness to items on the landscape? An immediate response would be to assign to corpuses our best possible *estimation* of fitness at the time. As in the worry above, going from the actual value to the best estimate of that value reduces predictive and explanatory power, but does not completely undermine the usefulness of the model.

There is strong similarity here with the Lenman/Burch-Brown argument presented above, which is why this discussion is presented here rather than in the last chapter. Given past experience with the downstream effects of the inclusion of new information in corpuses, practitioners become adept, to an extent, at assigning new items of information to one of several familiar classes of

causal chains, allowing them to give a rough estimate of the beneficial potential of the new item of information given the current state of the research field and the current identities and state of the information consumers. In adopting this view I position myself closer to Polanyi (1962) and further from Gillies (2014), and specifically I diverge from Gillies' position that the potential benefits of *all* research projects are inherently unknowable *ex ante* (see Chapter 1).

The ability to estimate fitness will not, however, be accurate for all projects in the landscape. The more dissimilar (distant) a project is from past projects, less knowledge will be available to predict its contribution to well-being. This is included formally in the model in the next chapter, by introducing a second scalar field on the landscape to describe the "visibility" of projects, i.e. the predictability of their future counterfactual benefits given accumulated experience.

Another, more subtle, aspect of the estimation of fitness is that the inclusion of a new item in a corpus is not instantaneous, but rather it is a process. The ability to estimate the fitness of an information item is likely to increase significantly over the duration of this process, for example once it is clear which journal the information is going to appear in, how it is going to be phrased, etc. Thus, it can be expected that the model would yield more accurate descriptions if estimations of fitness are considered at the point of inclusion in the corpus. More on this in the next chapter.

A final point on estimated and actual fitness relates to fitness dynamics, and will thus appear multiple times in this chapter. When the consequences of a process result in the deformation of the information landscape, is it because the actual fitness of the corpus has changed, or because the best estimate of the fitness of the corpus has changed? First, it is important to note that there need not be one answer: sometimes the actual causal chain would be changed in ways we cannot yet know about, and so actual (or counterfactual) fitness will change while estimated fitness will stay the same; at other times the causal chain will remain intact, but elements outside the causal chain will make us change our estimation of long-term benefit.

However, while *in principle* actual and estimated fitness may change independently of each other, in most practical cases they would not. This is due to a bidirectional influence between actual/counterfactual and estimated fitness:

Actual/counterfactual fitness → Estimated fitness: Influence in this direction is trivial: whatever process triggered the change in the causal chain that would follow the inclusion of an information item in a corpus, if funding bodies know about it, they can change their estimate of fitness accordingly. As mentioned above, in the analysis given in this thesis funding bodies are assumed to have complete access to knowledge available at any given time, and so, under this idealising assumption, are highly likely to have knowledge about the process that is taking place to alter the causal chain.

Estimated fitness → Actual/counterfactual fitness: Influence in this direction is less trivial, but is none the less expected to take place. The reason for influence in this direction is that funding bodies, and other institutions and individuals that share knowledge with funding bodies or rely on funding bodies for knowledge, play an important part in the causal chain that determines the beneficial outcomes of information inclusion in public corpuses. For example, if funding bodies' estimation of the fitness of a certain health claim is suddenly increased, more funds are likely to be directed to the field, which will increase the chances of refinements being made to the claim and its actual fitness being increased. In addition,

the estimation of increased fitness is likely to be shared by other public institutions, such as governmental health organisations, which by their altered actions would change the causal effects of the information item.

There is theoretical interest in exploring the above points in more detail, but for the sake of brevity I will cut short the discussion here and instead highlight specific instances of the above in the historical examples given below and the analysis that follows. The upshot of the above is taken to be that despite deep theoretical differences between actual fitness, counterfactual fitness, and estimated fitness, influences between these concepts will, in most cases, correct for slight misapplication of the right category, making the use of the model justifiable in the context of institutional critique and design even if the "wrong" category is being used to structure the topology of the epistemic landscape.

4.1.3 Fitness dynamics under varying notions of well-being

As mentioned in the previous chapter, there are many different accounts of well-being, and they can be quite different from one another. This thesis brackets out the consideration of the appropriate notion of well-being for the context of science funding, and assumes the model user comes to use the models presented in this thesis with an appropriate notion of well-being, whatever that notion may be. This openness, however, makes the discussion of fitness dynamics somewhat tricky, as different causal processes may be judged according to some notions of well-being as having changed the fitness of research projects, while according to other notions of well-being no change in fitness has occurred. One example of this is in the division between subjective accounts of well-being, which will be very sensitive to the information individuals in the society have about their own condition, and objective accounts of well-being, which are not as sensitive to this consideration. The problem of assigning fitness changes under different accounts of well-being also affects the distinction between actual and estimated fitness, discussed above, as, for example, under some accounts of well-being an erroneous but comforting theory may contribute to actual well-being, while on other accounts this theory may only be perceived as contributing but in fact make no contribution to actual well-being.

The solution offered to this problem is similar to the one presented above regarding estimated/actual fitness: the causal chain of events surrounding funding decisions, research activity, and publication is messy and complex, and will in many cases cross boundaries between objective and subjective, estimated and actual. While not all cases of fitness dynamics described below will count under all notions of well-being as genuine cases of dynamics (perhaps they will be judged only as perceived changes, perhaps the notion of change will be rejected entirely), under most accounts of well-being relevant to the science funding context the majority of processes described in this chapter will have some effect on well-being, and the accumulated effect of numerous processes (even if they are only a subset of the processes discussed in this chapter) will suffice to undermine funding strategies that focus purely on optimisation, such as peer-review.

4.2 Historical references

Later sections provide a theoretical framework for classifying and analysing dynamical processes that change the fitness of research projects over time. When we look at historical episodes, it is overly optimistic to hope to find clear examples of each process: these processes often co-occur,

and inter-relate in complex ways. I therefore choose to start with brief summaries of a few historical episodes, each exemplifying *some* of the processes analysed in later sections. These later theoretical sections will then refer back to these examples, singling out a particular causal process of fitness-change.

The historical summaries presented below do not represent original research. Each is based on sources of secondary literature in the history of science and technology. The first summary, of the development of the laser gyroscope, is based on MacKenzie (1998, ch. 4). The second summary, of the discovery of DNA's double helix, is based on Allen (1975).[1] The third, of the regulation and risk assessments following the creation of recombinant DNA molecules, is based on Krimsky (1982); Wright (1994). When reading these summaries, keep in mind that they do not tell the *whole* story, as scope is limited. Rather, they highlight specific cases from the history of science and technology when the *fitness* of a piece of research, or a whole field, changed drastically and rapidly.

4.2.1 The Sagnac effect and the laser gyroscope

Around the end of the 19th century, and on until the early 20th, there was an active debate among physicists about the ether. The ether, postulated to be the medium through which electromagnetic radiation propagated, played a crucial role in Newtonian accounts of electromagnetism, yet was undetectable by any known measurement or experiment. The search for an experiment to detect the ether was thus theoretically motivated, and its famous high-point is the Michelson-Morley experiment, conducted in the 1880s, which showed the ether, or rather the "ether wind" produced by the rotation of the Earth, was not detected in an experimental setup which should have, according to Newtonian electromagnetic theory, made it detectible. Later on, this experiment was recast as the "crucial experiment" which allowed Einstein's theory of relativity to triumph over its Newtonian counterpart, but this revised history greatly simplifies the actual process of events.

Georges Sagnac, a French professor of physics, planned an experiment to detect the "ether wind" generated by rotation. Unlike the Michelson-Morley experiment, which used the rotation of the Earth as the source of "ether wind", in Sagnac's experiment the table itself, on which the mirrors and camera were placed, was rotated. His experiment was successful in detecting an effect: the interference pattern of two beams, sent in different directions around a circular path and then made to converge, was different when the setup was mechanically spun, and the magnitude of the effect was according to the theoretical prediction of an "ether wind". This result, obtained in 1913 and submitted in a paper to the Acadèmie des Sciences, was positively received in France as "having verified the theory of the ether." It is important to notice that in France it was not until the late 1950s that relativity was accepted in teaching, textbooks, and university programs. In comparison, in the Anglo-Saxon world relativity was already widely accepted by 1913, and it was quickly shown that the Sagnac effect could equally be accounted for by Einstein's general relativity, due to relativistic time-dilation. While some, including Michelson, tried to revive the ether theory by repeating Sagnac's experiment, their success had little effect on the ascendancy of Einstein's theory of relativity and the demise of the ether.

This might have been the end of the story of the Sagnac effect, were it not for the advances in quantum electronics in the late 1950s. These advances introduced a new device, the resonant

[1] For a more lively, but less historically accurate account of the same episode, see Watson and Stent (1980).

cavity, which generated a powerful electromagnetic wave of a specific frequency, related to the physical dimensions of the cavity. These cavities provided the foundation for various technological breakthroughs, first in the microwave region as key components in radar and maser systems, and later in the visible light region with the invention of the laser. Between 1959 and 1961, the hype around the maser and the laser led three independent researchers, each of them familiar with the Sagnac effect through the same textbook (R. W. Ditchburn's *Light*), to propose the use of resonant cavities in a device that would use their powerful emissions to measure rotation – an electromagnetic gyroscope. The military and industrial need for better gyroscopes was known to all three researchers, who were working in, or affiliated with, industrial corporations which produced mechanical gyroscopes for the US military and for the aviation industry. These industries, in turn, required more precise and durable gyroscopes due to increased speed and manoeuvrability of their vehicles: marine, aircraft and missiles.

All three individuals, and their respective teams, tried to obtain resources for the construction of a prototype, both from internal sources and from military and government R&D funds. Initially, applications to external funding were rejected, and it was the team at the Sperry Rand Corporation, which had access to internal resources, who first succeeded in constructing a working prototype by January 1963. Their prototype was not a mere rendition of the Sagnac experiment done with lasers, but relied on a different physical phenomenon in the superposition of the light beams at the detector, which greatly increased the sensitivity of the device, and in principle offered a simple way of digitising its output, thus making it more appealing in an increasingly-computerised industry.

However, the performance of the 1963 device was much poorer than the performance of mechanical gyroscopes of the time, and, after a short period of media hype around the newly-invented "laser gyro", followed a long period of development work, involving mostly engineering but also detailed theoretical physics, to make the device better than the existing alternatives. This work was done in several corporations throughout the 1960s and early 1970s, only resulting in a first commercial contract in 1972. Throughout this period belief in, and support of, this project increased and decreased dramatically, with internal and external sources of funding used at different times, and with some corporations giving up altogether part-way through the race.

Eventually, the initial commercial success of 1972 was followed by a landslide success with the inclusion of a laser gyroscope as the inertial navigation system of Boeing's new airliners, the 757 and the 767, in 1978. However, the market dominance achieved by the laser gyroscope over its mechanical alternative, first in the airline industry and later in military applications in planes and missiles, could not be explained by technological superiority alone, as the performance of laser gyroscopes and new models of mechanical gyroscopes were comparable throughout the period; instead, it is clear that the laser gyro enjoyed a "technology charisma" which the mechanical gyro lacked, whether through the media hype surrounding the laser gyro, or simply because it utilised what was conceived as more advanced science.

The historical episode presented above exemplifies a significant shift in importance of a research result, in this case that of the Sagnac effect, from a curious footnote to the story of the demise of the aether to its key role in the development of a commercially successful laser gyroscope. This shift can be broken down into several processes of fitness dynamics, which can be generalised, as discussed in the next section.

4.2.2 The discovery of DNA's double helix

Two threads of the story of the discovery of DNA's structure can be traced back to the 1860s: one begins with Gregor Mendel's published work on heredity of characteristics in crossbred strains of the common garden pea, the other with the discovery by Friedrich Miescher of nucleic acid, a hitherto unknown substance which is contained in cell nuclei. A third thread starts in 1912, with the invention of X-ray crystallography by W. H. Bragg and W. L. Bragg, which allowed the experimental elucidation of molecular structures of large molecules, and the foundation of a British school of crystallographers.

The genetic thread of Mendel's work was picked up in 1900, and started a line of experimental work in genetics, which included the discovery that genes are arranged in a linear order on the chromosomes, and that genes were susceptible to mutations. These discoveries sparked interest among groups of physicists, who became interested in the chemical and physical puzzle of the material manifestation of the gene, and possibly undiscovered chemical and physical processes involved in it. In 1935, Max Delbrück, a pupil of Nils Bohr, published a speculative paper on the structure of the gene, postulating its role as an information carrier. This idea was picked up and popularised in 1945 by the famous physicist Erwin Schrödinger, in his book *What Is Life?*. In the mean time, Delbrück began experiments on gene transfer in the context of bacterial viruses, known as phages. In 1940, Delbrück, together with Salvador Luria and Alfred Hershey, started the Phage Group, with the explicit purpose of solving the mystery of the nature of the gene.

The biochemical thread of Miescher's work was continued, and by the early 1920s it was known that there were two kinds of nucleic acids: ribonucleic acid (RNA), and deoxyribonucleic acid (DNA). By the late 1920s it was known that DNA was located predominantly in the cell nucleus, whereas RNA was located mainly in the cytoplasm. Since the chromosomes were also located in the nucleus, this suggested a greater importance for DNA in the process of heredity. However, the chromosomes are made up of both proteins and DNA, and the consensus opinion was that genes were probably related to proteins, with DNA playing a secondary role. Part of this belief was based on the smaller number of basic components that make up DNA, only four nucleotides, as opposed to the 21 different amino acids that make up proteins. It was believed at the time that the nucleotides repeated in a simple pattern to form DNA. Advances in the crystallography of proteins in the 1930s (described below) demonstrated that the sequence of amino acids in proteins was non-repeating, making proteins the only known biological polymer of non-repeating structure at the time.

In 1944, Oswald T. Avery, Colin MacLeod and Maclyn McCarthy published a paper providing the first direct demonstration that DNA was the molecular bearer of genetic information. In a transfer of purified DNA from a normal donor bacterium to an abnormal recipient bacterium, the recipient bacterium transformed into the normal state, and descendants of the recipient also inherited the change brought on by the transferred DNA. However, on the background of the known biochemistry detailed above, Avery, MacLeod and McCarthy phrased their findings very cautiously, leaving open the possibility that "some other substance" has escaped detection or that DNA's role was only confined to the few traits tracked in the experiment. The reception of the results was very hesitant, and though wildly circulated, it was not accepted into consensus opinion about heredity.

However, in the late 1940s and early 1950s Erwin Chargaff produced experimental evidence that the relative amount of DNA nucleotides differed between species. Chargaff further showed

that pairs of nucleotides, adenine (A) and thymine (T) on one hand, cytosine (C) and guanine (G) on the other, appeared in almost identical concentrations, whereas the relative concentrations of AT to CG differed. On the backdrop of this changed biochemical background, a refinement of the experiment of Avery, MacLeod and McCarthy was conducted in 1952 at the Phage Group, by Alfred Hershey and Martha Chase. Their experiment showed that when phages infect bacterial cells, it is only the DNA of the phage that actually enters the cell. By using radioactive isotopes for tagging molecules, Hershey and Chase could prove that no protein was carried over with the DNA. This further evidence of DNA's role in transmitting genetic information, and the biochemistry that opened room for it to play this role, was sufficient to influence consensus opinion, and focus genetics research on DNA.

Throughout these episodes, the structural thread, launched by the invention of X-ray crystallography, was keeping pace. In the late 1930s, Bragg's pupils W. T. Astbury and J. D. Bernal (the same Bernal mentioned in §1.5) began to tackle the structural analysis of proteins and nucleic acids. Their work was followed by Max Perutz and John Kendrew, who worked on the structure of two oxygen-carrying proteins. A significant breakthrough in the field was achieved in 1950, when Linus Pauling predicted, and then confirmed, the basic structure of the protein molecule, dominated by a spatial feature called the α-helix. Pauling's work differed from the purely crystallographic work of the Bragg school, as it relied on a combination of crystallography, chemical insight, and model building.

The increasing interest in DNA, detailed above, led several groups to attempt to decipher its molecular structure. These groups included Pauling in Cal Tech, Maurice Wilkins and Rosalind Franklin in King's College, London, and Francis Crick and James Watson in Cambridge. Wilkins and Franklin have been pursuing the structure of DNA even prior to Chargaff's results, and their efforts were focused on purification and crystallography of DNA. Pauling was attempting to apply his chemically-informed model building method to DNA, but he had no access to the unpublished results of Wilkins and Franklin. Watson and Crick focused on model-building, driven by chemical and *biological* insight, and also had access to unpublished results by Franklin and Wilkins. In 1953 Watson and Crick published the now-famous paper in *Nature*, in which they describe the double-helix structure of DNA, and suggest its direct role in supporting life by offering a mechanism for replication. Watson and Crick's result had immediate and dramatic effect, and in the following decade was incorporated, through theoretical and experimental work, into what is now known as the central dogma of molecular biology.

The story of the discovery of DNA's double helix structure exemplifies several processes of dynamic fitness, perhaps the most striking of which is the effect of Chargaff's results on the impact of research done by the Phage Group, as seen in the contrast between the reception of the Avery, MacLeod and McCarthy paper and the reception of the Hershey and Chase paper (though by no means was Chagraff's result the only cause of difference in the reception of these papers). The various processes of fitness dynamics that took place in this historical episode are explored and generalised in the next section.

4.2.3 Recombinant DNA, regulation and risk assessment

The first successful experiments in artificially creating recombinant DNA (rDNA), a DNA molecule that contains segments of genetic material from two different organisms, took place in 1971. The Berg group at Stanford University Medical School successfully combined strands of

DNA from two different viruses, SV40 and lambda, and used the hybrid rDNA molecule to infect the bacterium *Escherichia coli* (*E. coli*). This success opened the possibility of a new tool, both in science and technology, of manipulating the characteristics of organisms well beyond what was previously possible via mutation and crossbreeding. However, this new tool also introduced considerable risks, in its capacity to create novel organisms with adverse health or environmental effects. Specifically, the SV40 virus was known to promote the creation of tumours in monkeys, and *E. coli* was known to reside in the human gut. The conceived risk was that an *E. coli*, infected with SV40 via the lambda virus as a vector, would become carcinogenic and spread into the human population through the laboratory workers. This process could only take place in the context of rDNA, because SV40 is a mammalian virus and does not, on its own, infect bacteria, whereas lambda is a phage, a virus that infects bacteria.

Theoretical, abstract knowledge of these risks was already present in the late 1960s, but the success of the Berg group made them urgent. The head of the group, Paul Berg, decided to suspend the experiment in 1972 until further information about the risk and its mitigation could be obtained. In the meanwhile, less risky experiments were carried out in 1972 and 1973, refining and simplifying the technique of creating rDNA, and showing that rDNA molecules could be made with genetic material from more complex organisms than viruses. In 1973, at the Gordon conference, a prestigious meeting of leading scientists that allows informal exchange of results and ideas, Berg raised his worries about the risks of rDNA experiments. These were picked up by the conference attendants and led to the chairs, Maxine Singer and Dieter Soll, sending an open letter to the National Academy of Science (NAS) to investigate the issue of rDNA and provide guidelines for safety.

NAS formed a committee to investigate the issue, headed by Berg. Berg invited a number of leading molecular biologists and biochemists, as well as Richard Roblin, a microbiologist with an interest in bioethics, to a meeting in MIT in 1974. The Berg committee discussed the risks involved in dealing with animal viruses in rDNA experiments, as well as more general risks, and considered possible containment and safety solutions that will enable carrying out these experiments. The committee's conclusions were drafted in an open letter which was sent to NAS and to the National Institutes of Health (NIH), the funding body for biomedical research in the US (see Chapter 1). In the letter, Berg called for a voluntary moratorium, or deferment, of certain kinds of high-risk experiments, until proper safety measures could be formulated. It also called for NIH to arrange an international conference to study the topic of rDNA and biohazard, and to form a regulatory body which will set the guidelines for safety measures and oversee (and fund) their implementation.

NIH responded favourably to the Berg letter, and between 1974 and 1975 it arranged an international conference, to be held at the Asilomar Conference Center in Pacific Grove, California, and officially formed the NIH Recombinant DNA Molecule Program Advisory Committee, which was later shortened to the Recombinant DNA Advisory Committee (RAC). At Asilomar, the 150 participants discussed the technical details of the experimental procedures of working with rDNA, the possible biohazard risks, and possible ways of mitigating these risks. It was suggested that experiments would be classified into categories based on the level of risk, and appropriate methods of containment be assigned to the different levels. The containment will take two forms: physical containment, similar to that practiced in laboratories working with pathogens, and biological containment, which will rely on using modified bacterial hosts (called *E. coli* K12)

which would not survive outside the lab. The technique for creating these weakened strains did not exist at the time, and it was suggested that resources be diverted to its development. The recommendations from the Asilomar conference were, by and large, endorsed by RAC, and formed the basis for the guidelines published by RAC in 1976. The publication of the guidelines effectively lifted the moratorium on almost all experiments, given the right containment measures were used.

The consensus opinion of Asilomar, reflected in the RAC guidelines, did not go unchallenged. There existed widespread agreement, both at Asilomar, RAC, and the wider community, that the risks discussed were largely unknown. While the Asilomar response to these risks was to define certain safeguards, these were criticised by some as being too restrictive, and by others as being insufficient because their effect at mitigating unknown risks was, essentially, unknown. The evaluation of these risks became the topic of three scientific meetings: in Bethesda, Maryland in 1976, in Falmouth, Massachusetts in 1977, and in Ascot, England in 1978. Unlike the Gordon conference, the Berg committee, and the Asilomar conference, which drew mostly on the expertise of molecular biologists and biochemists, these meetings also drew on the expertise of virologists, epidemiologists, and infectious disease experts. The consensus opinion which emerged from these meetings suggested that the risks were smaller than originally considered, and that the RAC regulations could be eased. While encouraging further studies into the epidemiological and ecological effects of rDNA, they recommended deregulation of research. These recommendations played an important role in further easing of the RAC guidelines, while in the background of the risk debate the number of rDNA research projects was booming, and results were getting closer and closer to commercial applications.

The story of the regulation of rDNA research exemplifies processes of fitness change that arise following breakthroughs in risky avenues, for example the greatly increased importance of novel solutions that can help in risk mitigation, such as the creation of *E. coli* K12. The various processes are explored, and generalised, in the next section.

4.3 Influences within a topic

The three historical examples presented above feature episodes in which the epistemic fitness of various research projects changed significantly. To give just a few examples, the fitness of research on the Sagnac effect just prior to the invention of the laser was much lower than its fitness just following the invention; the fitness of work on DNA structure just prior to Chargaff's work on relative nucleotide amounts was much lower than its fitness just after the discovery; the fitness of research on modified bacterial hosts, such as *E. Coli* K12, was much lower prior to the successful creation of rDNA and the recognition of potential hazards than its fitness following these events. This section and the next two sections present a classification of different causal processes by which the fitness of a project can change over time, as a consequence of research and development activity in the same or related fields. The three sections group different processes together according to the kind of interaction between the research and development activity causing the change and the research project affected.

This section presents an overview of all processes where the activity causing the change and the affected project both belong to the same "topic". In Weisberg and Muldoon's model (discussed in §2.4), a "topic" relates to the area of knowledge covered by, e.g., an advanced

scientific manuscript. To this "topic" Weisberg and Muldoon associate a landscape, which covers all the different *approaches* to researching this topic. In the "information landscape" model (§3.4), an information landscape can be associated with roughly the same scope of "topic", though this landscape is associated with the information contained in the associated corpus, and therefore covers the different *results* that could be obtained in the topic. A revised epistemic landscape (§3.5) was then defined as including all projects whose results belong to this corpus. Both Weisberg and Muldoon's model and my own revised epistemic landscape associate the height of each coordinate with a measure relating to results: in Weisberg and Muldoon's case this measure is *significance*, in my case it is *epistemic fitness differential* (§3.2.3). From Weisberg and Muldoon's model it is clear that there are many different groups of scientists pursuing different approaches in the same topic *simultaneously*. This section lists the various ways in which advances in one project, and consequently the adoption of its results into the relevant corpus, can influence the epistemic fitness differential of other projects in the same landscape.

The first three processes presented in this chapter are accompanied by visualisations of the effect using a top-down view of a three-dimensional version of the epistemic landscape. These visualisations serve a dual purpose: they help show the connection between the model developed in the previous chapter and the theoretical classification presented in this chapter, and they introduce a particular representation that will be used extensively in the next chapter where dynamical processes are combined using a computer simulation.

4.3.1 Duplication and redundancy

In many episodes in the history of science we can see several groups of researchers pursuing the *same goal*, often even using the *same approach*. In the historical summaries, we have seen several groups racing to be the first to demonstrate the Sagnac effect in a compact system using lasers; we have also seen two groups, Pauling's group and Watson and Crick, racing to decipher the structure of DNA using model building.

We have also seen, in Chapter 2, this scenario treated theoretically: this is the main scenario in the project-choice models of Kitcher (1990, 1993) and Strevens (2003). Strevens claims this race to the same result is the explanation for the specific "winner takes it all" reward structure in science, and the resulting importance of priority disputes. It is easy to see how, in the idealised case where all groups try to solve *exactly the same problem*, the success of the first group in reaching a result that is accepted by the community would make all other groups' efforts redundant; the first group's results would be incorporated in the corpus, thus bringing about any contribution to well-being that the result could causally support. Any later project that reaches the same result will have nothing new to contribute to the corpus, and so further increases in epistemic fitness would be negligible, if not outright zero, depending on the extent of the identity between the new result and its already-included counterpart.

In the epistemic landscape model, this dynamic effect can be depicted as follows: several scientists (or groups of scientists) occupy the same coordinate simultaneously, pursuing the same project in the aim of achieving this project's fitness-increase potential. Once the first group succeeds in completing the project and publishes the associated results there is no longer any benefit from further pursuing the same project, and so the coordinate's fitness differential (its height in the model) is set to zero, regardless of its value prior to the first successful discovery. This can be visualised using a top-down view of a three-dimensional landscape, as in Fig. 4.1.

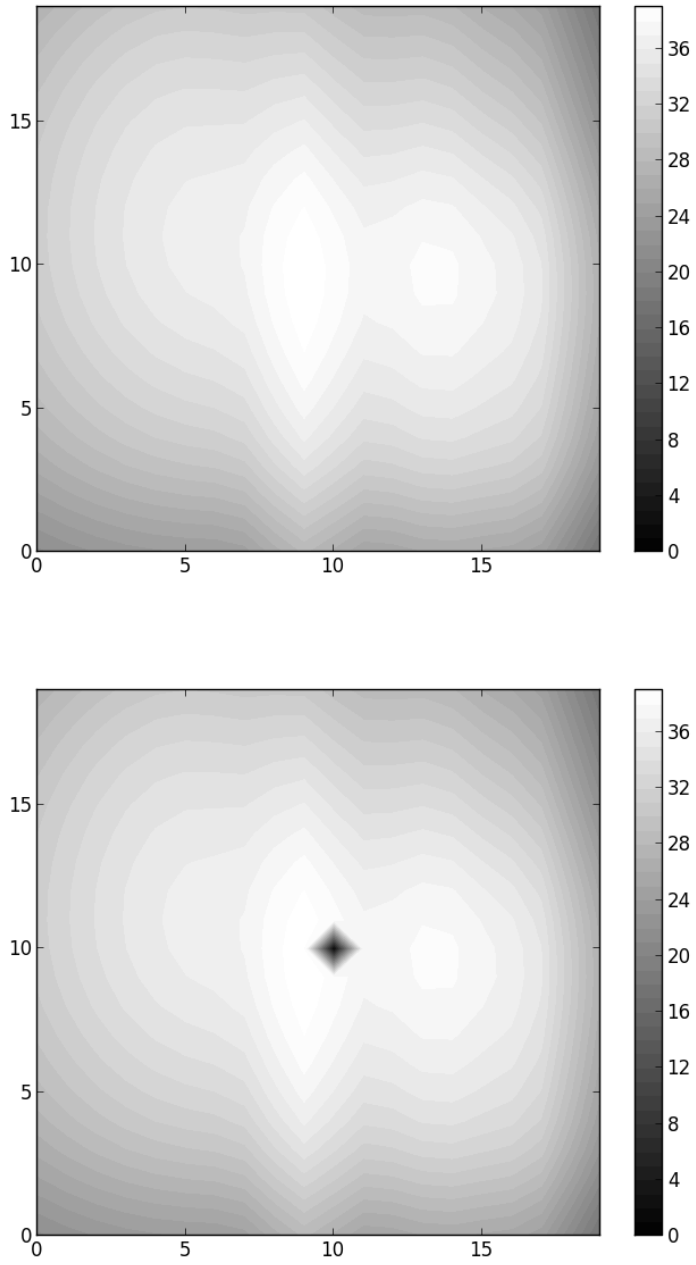


Figure 4.1: Visualisation of the "duplication and redundancy" process, where the fitness differential of a project is set to zero once it has been successfully pursued by an investigator or group of investigators. The top image shows a top-down filled contour plot of a hypothetical epistemic landscape prior to any exploration, while the bottom image shows the same landscape after research has been carried out on the project at coordinate (10,10), as a result of which its fitness was set to zero. The colour of each coordinate represents its associated epistemic fitness gain if the projects is pursued, according to the colour-bar on the right, in arbitrary units of fitness gain. The code for simulating and visualising hypothetical epistemic landscapes is presented in the appendix.

The effect is fairly localised, as it only affects one topic, and within that topic only the area of the landscape immediately near the relevant project. The magnitude of the effect, however, is very large, as it can overnight change the fitness of a result from very high (ground-breaking discovery) to almost zero (duplicate of known result).

4.3.2 Convergence: novelty versus support

A similar scenario to the one described above, but perhaps more realistic, is when groups pursue *different projects* that nonetheless converge on the *same result*. In this case there may be small fitness differences between the results, due to extra information about, e.g., the evidential support for the result. Nonetheless, it is clear that the first group who successfully leads to the inclusion of the result in the corpus will have the highest influence on the fitness of the corpus, effectively reducing the fitness contributions of all other groups working on similar projects (though not to zero).

The clearest example is the relation between Watson and Crick's result and the projects of the X-ray crystallographers. By pursuing a different approach they reached the fitness peak (structure of DNA) first, thus reducing the fitness for all other projects on the structure of DNA. However, their result was based on model building and chemical and biological insight, and required confirmation from structural data, i.e. from X-ray crystallography. Thus, the results of Wilkins and Franklin were not of zero-fitness, though they clearly had much less influence on the eventual chain of events than in the counterfactual history where Watson and Crick did not exist, and the double helix structure was discovered purely using X-ray crystallography data.

In the terms of the model, we can see this process of convergent results as an extension of the effect from the previous section. Once a group completes a project of sufficient importance and novelty (a project which is assigned a fitness contribution above a certain threshold), nearby (similar) projects lose some (though not all) of their fitness because some of the beneficial causal consequences that would have accrued from their pursuit had already been gained by the project that has just finished. Further work may build on the original contribution and thus generate further useful outcomes, but the gain in fitness will be lower than the original fitness gain. This process can be visualised in a similar manner to the process described in the previous section, as in Fig. 4.2.

In addition to significant fitness-gain reduction on the side of the hill where the major discovery takes place, there is also more limited reduction of fitness on other sides of the hill, as results change fitness from very high (ground-breaking discovery) to medium-low (novel evidence for existing result). The effect is more widespread than in the previous section, but smaller in magnitude.

4.3.3 New avenues

Not all avenues of research are conceivable at all times, nor the full picture of how the results of research could be used to promote well-being. If an innovative/unexpected line of research succeeds, it opens the way for much new research to follow, revealing the fitness potential of its descendants and mapping out new previously-unknown parts of the landscape.

The establishment of DNA as the carrier of genetic information by Chargaff and Hershey and Chase greatly increased the fitness of all DNA-related research. More specifically, Hershey and

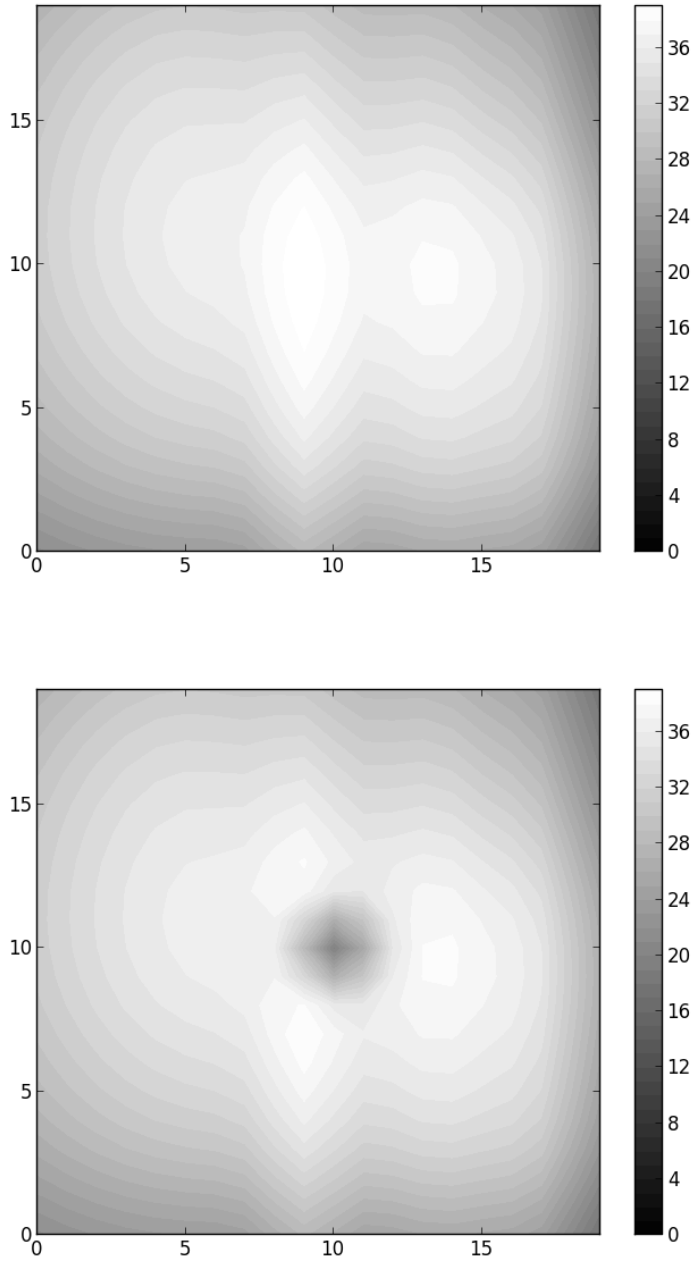


Figure 4.2: Visualisation of the "convergence" process, where the fitness differential of a project and its surrounding (similar) projects is reduced, though not to zero, once it has been successfully pursued by an investigator or group of investigators. The top image shows a top-down filled contour plot of the epistemic landscape prior to any exploration, while the bottom image shows the same landscape after research has been carried out on the project at coordinate (10,10).

Chase's result opened up new avenues in researching DNA using phages, and Chargaff's results greatly increased the viability of Watson and Crick's modelling approach. The effect of Watson and Crick's result was even greater: immediately, many new experiments became conceivable, and their relevance apparent. These experiments laid the evidential basis for the "central dogma" of molecular biology.

In the epistemic landscape model, this effect can be represented as the emergence of new hills, triggered when a project of sufficiently high fitness-gain is successfully completed (see Fig. 4.3). While the magnitude of the effect of Watson and Crick's discovery is exceptional, smaller instances of this effect constantly happen in research; the results of a successful experiment almost always lead to new ideas and new experiments, adding fitness to locations near the coordinate of the approach that has been explored. We can assume that the probability of generating new hills, and the range of effect in which new hills can appear, will be positively correlated with the fitness of the discovery, and model accordingly.

4.3.4 Inertia

The fitness of research is not determined only by its empirical adequacy, but also by the ability of users to use it, which in turn partially depends on the presentation of the results and familiarity with them. If a research program has been pursued for a long time, its users are likely to be familiar with some of the terminology and methods it deploys, increasing the fitness of descendants of that program. This leads to a form of inertia which keeps research programs going given that they have been historically pursued, and thus is a form of contingency in the path of science. This adherence to historically successful projects has been commented on by Popper, Kuhn, and Lakatos.

Since the future fitness of the Sagnac experiment as a basis for a laser gyroscope could not have possibly been foreseen at the time, the actual fitness of the experiment in 1913 can most readily be explained by the inertial fitness of the ether theory, which kept it alive long after the conflict with the Michelson and Morley experiment and after an alternative, Einstein's theory of relativity, was offered. The inertial effect also explains the languishing fitness, for several years following the publication of Avery's results, of projects which assumed proteins were the bearers of genetic information. Inertia influences not only estimated fitness, but also actual fitness: French scientists were very happy to see results supporting the ether theory, and rewarded Sagnac accordingly, and sustained work on proteins yielded useful empirical results which were later incorporated into the central dogma of molecular biology.

In the model, the existence of inertia can be modelled as damping of other processes: when, for any reason, the fitness of a tall hill is reduced, the inertial effect causes the reduction of fitness to take place over an extended period of time, rather than instantaneously. We can assume that the damping effect is positively correlated to the height of the peak prior to reduction, and model accordingly.

4.3.5 Revealed risk

Some natural and artificial phenomena can have adverse effects on well-being. When these phenomena are the subject of research, there can be cases where initial research raises worries that further research might bring about adverse effects, either because of interference with a

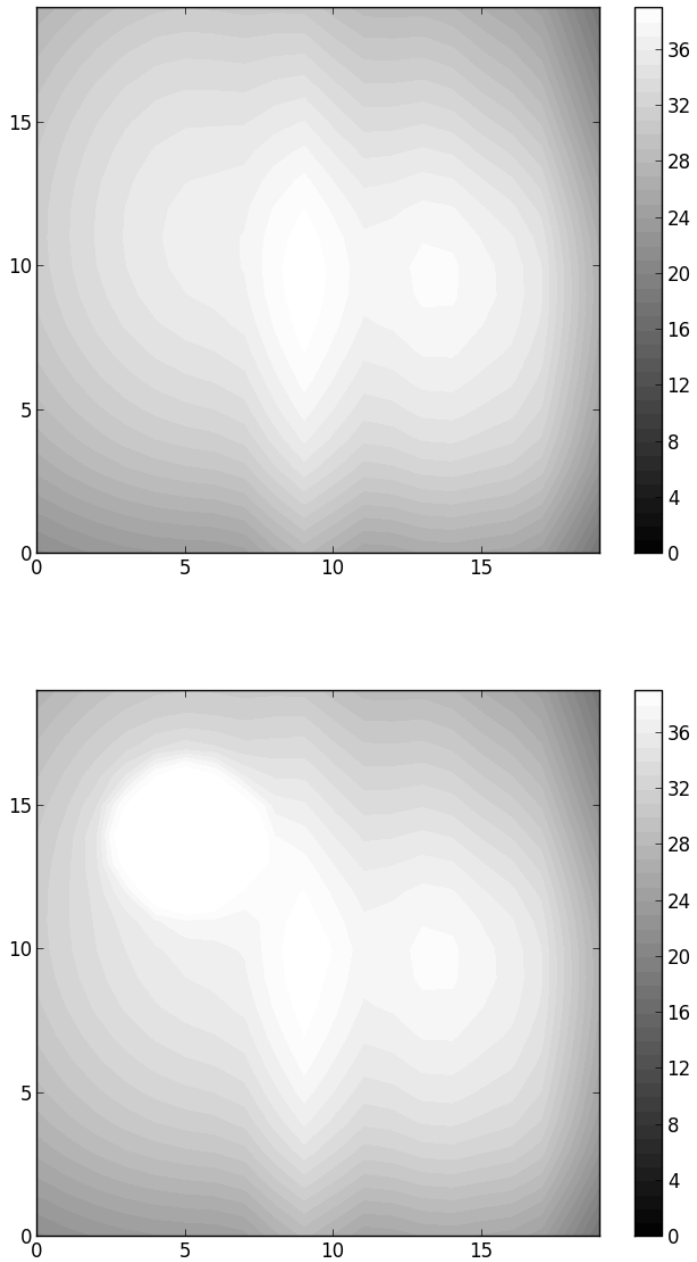


Figure 4.3: Visualisation of the "new avenues" process, where the fitness differential of a cluster of projects suddenly increases, once a novel and significant project has been successfully pursued by an investigator or group of investigators. The projects in the new cluster need not be similar to the project that has been completed, and so this effect is visualised as the appearance of a new hill somewhere on the landscape. The top image shows a top-down filled contour plot of the epistemic landscape prior to any exploration, while the bottom image shows the same landscape after research has been carried out on the project at coordinate (10,10). In this case the cluster of projects with increased fitness-gain is centred around coordinate (5,14).

natural environment or because its results could be used, either accidentally or maliciously, to cause harm. In such cases, the risks involved in further research can lead to the placement of restrictions on research or to more hesitant adoption of research from the field, which affect the causal chains of further new discoveries and alters their actual fitness.

An example of this process is the moratorium on certain kinds of rDNA research, which was put in place following initial successes by the Berg group in combining lambda phage DNA (which can infect bacteria found in humans) with SV40 viral DNA (which can cause tumours in monkeys, and possibly humans). The moratorium extended beyond Berg's own experiments, to all experiments of its kind that contained similar risks, thus effectively reducing the fitness of rDNA research, at least until the moratorium was lifted.

In the epistemic landscape model, this process can be represented as a small chance that, when researchers pass a relatively low threshold of a new hill, the height of coordinates higher up the hill is reduced. The duration of the effect can be permanent (completely neglected lines of investigation due to risk), but more often the reduction of fitness is temporary, until safety measures can be put in place.

4.4 Influences between topics

The previous section described how different projects on the same topic can influence each other's fitness. These processes could, in principle, be represented in Weisberg and Muldoon's model (though they are not included in Weisberg and Muldoon's simulations), because they all occur within the same landscape. However, we also need to consider interaction between research projects in different topics. For this we need a picture that shows the link between different topics, which is provided by the models presented in the previous chapter. The previous chapter has shown that the scope of a landscape can range from very restricted (the information in a single paper) to very broad (all public scientific information). Thus, we are free to label all processes as occurring "within" a very broad landscape, or "between" very narrow landscapes. The distinction presented here follows Weisberg and Muldoon's choice to limit the scope of a landscape to the breadth of knowledge covered by a single advanced manuscript, as it represents a convenient level of analysis.

4.4.1 Reduction and emergence

Different topics could be related by virtue of relating to the same "chunk" of reality, albeit at very different levels of abstraction/analysis/ontology and with very different theories/practices/methods. One well studied instance of this relationship is a relationship via reduction and emergence, or a vertical ontological relation (Nagel, 1961). When such a relationship is established or explored, it can make some research projects and/or results redundant or obsolete, while opening up new ones. In some extreme cases these interactions can open up new interdisciplinary topics/landscapes.

If we look at the background to the discovery of DNA, we can see that the link between heredity and chemistry was not always there. It was formed first as a working hypothesis, by Delbrück and others, and then tested, or developed experimentally, by the Phage Group.[2]

[2]This is a simplification of the actual history of the field. For a more nuanced account see the details in Allen (1975).

The choice of the Phage Group to work on the physically smallest process of genetics, that of bacteriophages infecting bacteria, was a direct result of their desire to understand heredity and genetics as close as possible to the underlying level of chemistry and physics. Their success paved the way to the links between genetics, chemistry, and crystallography, which enabled the initial successes, and then boom, of molecular biology.

In the model, we can think of reductive successes as forming new links between previously unlinked topics. These successes, which usually originate within one particular topic, have widespread effect: they increase the fitness of all projects in the linked topics which can benefit from inter-topic integration, and they can form entirely new topics (landscapes), with new hills to explore. Of course, the opening of whole new topics is rare, but more local reductive successes can still generate new hills in their own, and in the linked, topics. Again, we can assume that the greater the fitness of the discovery that led to the reduction, the more widespread the range and magnitude of new hill formation, and the greater the probability of creating a new topic.

4.4.2 Practices and technology

Scientific practice relies heavily on available artefacts, both physical (equipment, synthetic reagents) and procedural (instructions, timings). The equipment and techniques used in one field are often the result of research in another field. Thus, the development of new tools can influence the fitness of projects that deploy these tools, increasing the relative fitness of those projects that can make use of the new tools compared to those that cannot.

The invention of the laser had far-reaching effects in enabling new kinds of research. One such example is the effect it had on the Sagnac effect, turning it from a historical curiosity about the ether to an active field of research and development in the market of inertial measurement. The specific details of the Sagnac effect were of no particular interest (low fitness) in the period between 1920 and 1960, until these became highly relevant (high fitness) to the development of new technology made available by the invention of the laser.

In the model, the invention of new technology and its effects can be represented by the transformation of old explored projects, with their fitness diminished following successful exploration, rising again to high fitness levels, following a successful discovery in a related field that offers new techniques. We cannot often know in advance whether the exploration of a peak will result in new techniques, but we can assume the probability of new techniques being generated is proportional to the fitness of the discovery. The notion of relatedness here depends on shared or overlapping methods and causal processes: the discoveries in quantum optics were linked to the Sagnac effect because the latter involved the manipulation of directed light beams.

4.4.3 Environmental effects and risk assessment

Sometimes, a basic research project translates into a new technology, which holds promise to increase well-being. However, due to its novelty, the implementation of the technology contains unknown risks, which create a demand for information about these risks. Thus, the success, or even the promise of success, of research in one topic, can directly increase the fitness of research projects in other areas, which, depending on the complexity of the environment and the size of the expected effect of the technology, could go well beyond the area of the original research.

The successful creation of rDNA molecules turned the risks of genetic engineering from an

abstract, vague worry to an urgent and specific need. The scientists involved with the research, and those familiar with it and tasked with regulating it, suddenly had an urgent need for information about the potential generation of biohazards and adverse environmental effects, which required expertise beyond that available in molecular biology and biochemistry. Thus, as a direct result of Berg's successful experiments with rDNA, the fitness of projects in epidemiology and virology that could estimate the risks, as well as projects in bacteriology that allowed the creation of safer hosts, rapidly increased.

This effect can be modelled in a similar way to the process in the previous section, except the new hills in the related topics will not necessarily be where the old hills were. The type of link between topics here is different though: topics are linked not only because of shared methods, but also because the causal processes they investigate occupy the same environment, either in the lab or in nature.

4.5 Influence of research on audiences

The previous two sections looked at how advances in some research projects can directly influence the fitness of other research projects, both in the same field and in other fields. This section looks at how advances in research can influence other research projects indirectly, via influence on the expected users and uses of information.

4.5.1 Communication technology

Many of the causal pathways from corpuses to information consumers rely on technology: cables and radio, internet protocols and digitised routers. These technologies are themselves dynamic, and change as a result of research. New disruptive information technologies can have drastic effects on the information consumption behaviours of individuals, groups, and societies, and drastically alter the readership of a particular corpus of information, potentially altering the fitness of the items of information contained within.

There is an interesting feature about all the historical episodes described which may at first escape the modern reader: they all took place prior to the invention of the internet and email, which means all information transfers had to take place by phone or letter, and all discussions by meetings and conferences. This is one reason why some of the events took years, rather than months, weeks, or days, to unfold.

It would be too hard to model the link between specific technological advances in computing and communication to the overall speeding-up and opening-up of communication. Rather, we can assume that overall the technological trend is towards faster communication, along greater distances, with larger audiences, for more contexts. Thus, all other processes of fitness dynamics discussed in this chapter will have a tendency to occur faster, and with a greater range of influence, as time passes.

4.5.2 Generating hype

When a research, or development, project is going to take time and money to reach fruition, as is often the case, researchers often find it useful to generate a media hype around early results, in the hope that this hype will provide immediate support, and also improve the reception of future results. This process can lead to a self-fulfilling prophecy, or a positive feedback loop, where the

successful reception of a product of research or a new technology owes much to the expectations generated by the initial hype that followed very early success.

When it won the 757 and 767 contracts, and came to dominate the market, the laser gyro was not, technically, vastly superior to its mechanical alternatives. However, it enjoyed a "technological charisma", which can be traced back ten years to the original media hype around the first successful prototype. This hype brought the existence of the Sagnac effect, and the promise of a commercial laser gyro, to the minds of prospective clients, in the military and aviation industries.

In the model, we can represent the effect of hype as the generation of new audiences to an existing corpus, following an initial exploration of a project with high fitness gain. The fitness of projects in the landscape then needs to be updated given the new audience, and in the case of hype this will take the form of increasing the height of coordinates which are near the initial successful discovery.

This is one of the cases where the specific notion of well-being used to evaluate projects' contributions makes a big difference, as discussed in §4.1.3. On subjective accounts of well-being hype can have a significant effect, whereas on objective accounts it may have little or no effect. As mentioned earlier in this chapter, this is complicated by the causal effects of subjective assessments on actual reception and related actions, which may end up having a real effect on well-being.

4.5.3 Isolation and boundaries

The converse of the trend towards faster and more open communication is a trend towards increased secrecy and control. Sometimes, meteoric and unexpected success of a research field can have adverse effects in terms of access to the information. This process can take many forms: a discovery can have relevance to national security, in which case it becomes classified; or it has promising economic potential, in which case it becomes protected by patents or kept as a trade secret; or it might be appropriated by more established fields of research, in which case it may be protected by partial secrecy or buried in jargon; or it may confer benefits to anyone who can practice a new set of skills, in which case formal bodies of accreditation are formed. In all of these cases, the transfer of information out of the field becomes more difficult, limiting the audience of the research and affecting its fitness.

During the first years of the debate about risks and regulation of rDNA research, the range of discussants was deliberately kept limited to scientists in molecular biology and biochemistry, who had an appreciation of, and an interest in, promoting the field of rDNA research. This gate-keeping took many forms: in hand-picking the individuals on the Berg committee, in using technical language in publications in *Nature* and *Science* that covered the debate, in limiting the number of participants and in providing a narrow framing of questions at Asilomar, and in inviting specific participants to Bethesda and Falmouth rather then publicising the conferences in academic journals as is customary.

The process of access restriction can be modelled as a reduction in the number of information consumers associated with the field in which a novel discovery has been made. The fitness for coordinates in the associated landscape then needs to be reevaluated given the new, narrower, audience, and we can expect that overall fitness of projects in the topic will decrease, with areas directly adjacent to the discovery losing more fitness than those further from it. Furthermore,

we can assume that the degree of restriction, and the associated degree of fitness reduction, will be correlated to the fitness of the discovery made.

4.6 The wider importance of fitness dynamics

The analysis of dynamic epistemic fitness presented above carries relevance beyond the scope of a theoretical consideration of optimal funding allocation. Directly, the introduction of dynamics to fitness, and similar measures of epistemic good, brings into question results in the theory of the social organisation of science that rely on static versions of these measures. A survey of works impacted by the introduction of fitness dynamics is presented below.

The analysis of fitness dynamics presented above is also an example of the use of models to bridge historical examples and theory development in the context of social design. Some comments on the use of models in this context are presented below, relying on the discussion of model-based strategies presented in previous chapters.

4.6.1 Arguments assuming static fitness

Funding allocation is not the only aspect of the social organisation of science that takes into account the fitness (or significance, or utility, or merit) of individual scientific projects. Works from across the field have used the notion of scientific merit (using one of the terms mentioned above) to further various arguments, and all have relied on merit being fixed while something else, often the distribution of investigators, varies over time.

Peirce, Kitcher, Strevens, Weisberg and Muldoon

One important locus of the use of static fitness is in the literature on the division of cognitive labour, surveyed in Chapter 2.

Peirce (1879/1967) used utility functions to obtain an equation for the optimal distribution of resources between alternative research projects, concluding that early investigations of new fields should be prioritised over further investigations of already-explored fields. Peirce's model alludes to a particular form of utility dynamics: the overall utility in a field will decrease over time, as more research is conducted on projects in the field. As shown above, this is but one of many processes through which the benefit of a research project can change, some of which have exactly the opposite effect: increased investment in a field can unlock hidden potential or increase its relevance to society, which will make future investment more worthwhile.

Kitcher (1993) used static utilities to argue that scientists who are motivated by the pursuit of credit can perform just as well, if not better than, scientists who are only motivated by a search for truth. Strevens (2003) used static utilities in his development of Peirce's and Kitcher's work, to argue that a specific credit mechanism, the scientific "priority rule" which operates on a "winner takes it all" basis, achieves optimal allocation of cognitive labour. Finally, Weisberg and Muldoon (2009) used static significance to argue that a mixed community of interacting scientists, specifically of "followers" (who explore regions of the epistemic landscape close to other investigators) and "mavericks" (who explore regions of the epistemic landscape far from other investigators) performs better than a community of non-interacting scientists (who explore regions of the epistemic landscape only based on local significance values).

By neglecting fitness dynamics, the importance of the decisions of the cohort of researchers for the generation of beneficial outcomes is artificially amplified in the models above. It is very possible, though not explored further here, that in the presence of fitness dynamics the motivations or characters of individual scientists, and their resulting actions, will have a rather small effect on the overall fitness gain achieved over time, and thus the explanatory power of these models in explaining social phenomena in science will be diminished.

Zollman, Hegselmann, Grim

Project choice is not the only field of research that relies on fixed merit.

Zollman (2010) used static benefits of alternative approaches to argue that optimal choice of tools/strategies/theories in science occurs when the community adopts transient diversity, i.e. when the community first adopts a diverse approach, with agents rapidly changing between, or distributing among, many alternatives, but then slowly converge on the most successful alternatives as their advantages become apparent. In the presence of fitness dynamics, the identity of the most successful alternatives will likely change over time, and so an optimal strategy will not be one of convergence, but rather one of continually managing transitions in the most efficient manner, alternating between states of convergence and divergence in complex patterns.

Hegselmann and Krause (2009) used a highly simplified epistemic landscape, a δ-function that represents "truth" on a linear opinion line, to explore the dynamics of opinions in an interacting population of mixed investigators and non-investigators. Hegselmann and Krause argue that we can be pretty optimistic about the capacity of the population to converge, in the long run, on the truth; the location of this truth is, in their model, static. While "truth", unlike epistemic fitness, may indeed be fixed, there is less comfort in a model showing that a society may converge on truth in some arbitrary matter over time, than there is in a society being able to converge on the truth in the matters of importance to that society. Which matters are of importance to a society, and which truths about them matter the most, are time-dependant targets which are amenable to the processes of fitness dynamics discussed above, and so Hegselmann and Krause's optimistic result cannot readily be extended to this more interesting case.

Grim (2009) used a static epistemic landscape to explore how various types of network structures *between* scientists promote or deter efficient solution of a difficult scientific problem, represented as a narrow peak in the landscape. Grim argues that too much connectivity can deter efficient exploration, as connected scientists are pulled towards broad, but lower, peaks. As in the case of Kitcher, Strevens, and Weisberg and Muldoon discussed above, the focus on the actions of scientists while neglecting the dynamics of the field itself may result in an artificial amplification of the importance of the actions being considered, and a correction of the amplification would reduce the explanatory power of Grim's model.

4.6.2 Fitness dynamics and models

The crux of the sceptical argument I develop based on the analysis presented above lies in identifying a large array of parameters that influence whether beneficial consequences will happen, and then arguing that agents are not in a position to know the values of these parameters to a sufficient level at the time of making their decisions. In the case of science funding decisions,

the parameters are the dynamic processes that will result in a change of fitness for funded projects. I argue for the existence of a situation I label a **"parameter crisis"**, where

The number of parameters required to make justified predictions exceed the capacity of an individual to satisfy them empirically.[3]

The argument in this chapter is presented to undermine the justification of predictions regarding specific fitness assignments to research projects, of the kind required to choose to fund one project rather than another. The argument may also apply, though this is not explored in this thesis, to more broad aspects of fitness assignment over long time periods, of the kind required to make a long-term funding strategy for broad fields of research.

It is, however, nearly impossible to *prove* the existence of a parameter crisis. To prove that, in a particular situation, the agents do not have sufficient knowledge to foresee the actual outcomes of alternative actions, it is necessary to show what knowledge *would* have been sufficient, or to show that no knowledge would have been sufficient. To prove either, it is necessary for the person making the evaluation to know more than the agents themselves, which in the context of science funding is unlikely to happen, except in the case of historical examples. This consideration motivates the focus on historical examples in this chapter. It was a similar consideration that motivated Gillies' use of historical examples in his critique of peer review, as discussed in §1.6.

The focus on historical examples, however, leaves open the possibility to argue against the validity of comparison between historical cases and current decision making, especially since funding bodies can claim they have learned from past mistakes. It may not be possible to counter this argument directly, but it is possible to argue, as I do in this chapter, that certain processes appearing in the historical examples display regularities which are generalisable, and we could therefore inductively expect them to apply, with some modification, both to historical and present contexts. The uncovering of generalisable regularities thus dictates the structure of the theoretical sections of this chapter.

A different objection to the sceptical argument is that it over-emphasises the importance of individual (regular) processes. Indeed, any of the processes presented in this chapter, taken in isolation, would usually not undermine the ability of agents to make choices between proposed research projects. This is why the sceptical argument is based on a "parameter crisis", the argument that, when combined, the processes paint a causal picture with too many parameters to know in advance. Since this causal picture is so complex (indeed, it needs to be for the argument to work) it is not possible to describe it directly. A full account of the dynamics of epistemic fitness, even only as it pertains to basic science, is well beyond the scope of this thesis. Luckily, it is not necessary to give a full account of fitness dynamics to support the sceptical worry. It is sufficient to show that we can assign some lower-bound to the complexity of the causal processes in the system, and therefore an upper-bound on its predictability (which will be contingent on the agent's ability to collect information). If this upper-bound lies below the level of predictability we would reasonably expect of the decision-maker, the sceptical argument holds.

The use of a model, or a combination of models, is particularly suitable for the task of setting a lower-bound on complexity. By their nature, models abstract away most of the complexity of the system, focusing on one, or few, causal processes (see discussion of models in Chapter

[3]The term "parameter crisis" is used in a similar fashion in the Geographical Information Systems (GIS) literature (Burrough, 1989; Turner, 1989, present early examples of its use) to describe models that have reached such a level of fidelity that for adequate use they require more empirical data than is available to the model-user.

2). Due to their open nature, models can then be combined to give a complex picture that integrates several causal processes. It is this feature that makes them useful in design, as in the example of design of auction systems discussed in §2.2. This feature also makes them useful in setting lower-bounds on complexity, as the complexity of the real system will be no less than the complexity of the union of all models that genuinely apply to that system, as each model is a minimally-complex representation of one part of the system. The only caveat to this line of reasoning is that the behaviour not represented in the model could have a countering or dampening effect, thus reducing the apparent complexity of the overall system as more causal processes are taken into account. However, I see no reason to think this is the case in the context of the dynamics of epistemic fitness, and thus conclude that the model-based approach, pursued in this chapter and the next, is well suited to argue for the existence of a parameter crisis in science funding decisions.

Conclusion

This chapter presented a list of dynamic processes by which advances in research can influence, directly or indirectly, the fitness gain potential of other research projects. The aim of this exercise is to make explicit the areas of knowledge required to make rational, reliable decisions about research project choice, *assuming* the aim of the deciding agent is to maximise eventual contribution to well-being. No claim is made that all of these processes occur all of the time, in all fields of research. Rather, I contend that these processes *could* occur at *any* time, in many different areas of research, and should therefore be taken under consideration when designing *any* rational policy of science organisation that aims to maximise future gains in terms of contribution to well-being.

The modelling work presented alongside the list is provided both as a way of clarifying the shape of the processes, and as a tool for combining the different processes together into one big picture of the dynamics of epistemic fitness. Given an instance of the epistemic landscape model, we could then specify the dynamic processes we wish to consider in that model system, with any combination of processes, from a single process to a combination of them all. By playing around with the parameters, we could then simulate the effects of the combined dynamics on the overall fitness changes, and from this we could estimate the capacity of evaluators to predict the future fitness of any coordinate on the landscape at any future time. In the next chapter I estimate some reasonable values for the various parameters, and provide a (rather pessimistic) evaluation of the scope for rational planning in science funding decision making.

Chapter 5

Simulating Funding Strategies

Random selection
can outperform peer review
in simulations.

Introduction

The previous two chapters have developed a conceptual framework for thinking about science funding mechanisms in the context of institutional critique and design. Chapter 3 presented a revised version of the "epistemic landscape" model as a way of representing the choices of funding bodies in a way that links these choices to the aim of funding bodies, namely an eventual increase in well-being. That chapter also introduced the notion of epistemic fitness as a measure of success for funding bodies' decisions, where epistemic fitness measures the extent to which the information stored in the public corpuses of a society contribute (causally) to an increase in well-being. Chapter 4 argued for the importance of changes in epistemic fitness over time to the evaluation of science funding strategies. That chapter also presented various sketches of how different processes of fitness dynamics may be represented using the epistemic landscape model.

This chapter builds on the previous chapters by constructing and presenting a computational simulation of alternative funding mechanisms. The simulation is based on the revised epistemic landscape model. The simulation includes quantitative versions of some of the dynamic processes explored in the previous chapter, though most are left outside the simulation for ease of calculation, and may be included in future work.

For initial validation, some cases for which we have strong intuitions are simulated and show a match between the simulation results and expectations. The simulation is then used to explore more complex cases for which our intuitions are not so strong. Specifically, the simulation is designed to compare alternative funding strategies in their ability to promote efficient exploration of fields of research and support the generation of information that will contribute to well-being. The simulation deliberately represents funding strategies' reliance on previous successes in their attempt to reach this positive outcome, and the conservative bias that this focus on previous successes may generate. In this, the simulations address the important tension between seeking novelty and seeking plausibility in the selection of new research projects, a tension that was highlighted in the contrast between Polanyi's and Gillies' arguments in the first chapter.

§5.1 revisits the model presented by Weisberg and Muldoon (2009), which was discussed

in §2.4, to show the mechanisms used in simulating a population of scientists exploring an epistemic landscape. §5.2 provides a high-level description of the simulation, its aim and its components. This high-level description is then expanded to a low-level detailed description of the simulation in §5.3. The latter section covers the different aspects of the simulation, including simulation components, chosen algorithmic representations, key parameters, and value ranges. The implementation of the simulation in code is presented in the appendix.

The results obtained, about the influence of fitness dynamics on the relative performance of various funding strategies, are presented in §5.4, and are both significant and surprising. The simulation shows that, under reasonable parameter values for at least some fields of science, choosing projects at random performs significantly better, in terms of fitness accumulation over time, compared to other funding strategies. The results support, to an extent, the proposal made by Gillies (2014) (discussed in §1.6), that science should be funded by random selection. These results, and their implications, are discussed further in the next chapter.

§5.5 and §5.6 outline possible criticisms of the simulation, and replies to these criticisms. §5.5 considers the epistemic value of a non-empirical simulation. §5.6 lists various unrealistic assumptions made in the model, and how they might be addressed in more advanced simulations. Specific worries about the interpretation of the results, rather than the mechanism used to obtain them, are discussed extensively in the next chapter.

5.1 A second look at Weisberg and Muldoon's simulation

The simulations in this chapter take place on a revised version of the epistemic landscape model, described in §3.5. The chosen representation and parameters for the simulation are influenced by previous work, most notably Weisberg and Muldoon (2009). This work has been discussed in broad terms in §2.4. In that chapter it has been argued that Weisberg and Muldoon's model, while offering significant advantages over previous models, still suffer from a narrow scope and restrictive assumptions that make their model, in its present form, unsuitable for analysing funding mechanisms in science. Nonetheless, the simulation *mechanism* developed by Weisberg and Muldoon, which was not discussed at length in Chapter 2, provides an important basis for the work presented in this chapter. The details of this mechanism, which involves a step-wise simulation of a population of scientist-agents on an three-dimensional epistemic landscape, are therefore presented in brief in this section as background for the model description given in later sections.

5.1.1 Aim of the simulation and key findings

Weisberg and Muldoon explore the microstructure of the division of cognitive labour, and how it is influenced by different types of individual research strategies. Specifically, they are interested in exploring the role of diversity (of approaches) in the distribution of cognitive labour. To study this, they investigate the relative success of populations of researchers in finding significant results, given population differences in the individual scientists' motivation to follow other scientists' approaches or explore novel approaches.

Using an agent-based population simulation on an epistemic landscape (described below), Weisberg and Muldoon show that the introduction of *mavericks*, individual scientists who are motivated to explore novel approaches, greatly increases the success of the population in finding

significant results. They show that when a single significant result is required, a pure population of mavericks, even a small one, will greatly outperform pure populations of followers or hill climbers, as well as mixed populations.

5.1.2 The epistemic landscape

Weisberg and Muldoon's simulation takes place on an epistemic landscape which is composed of a configuration space of *approaches* to which is assigned a scalar field of *significance*. They describe the landscape as a model of the different approaches a scientist could take to investigating a certain topic, where distance between approaches is given by their similarity: the more similar two approaches are, the closer they are. Each approach is assigned a significance value, representing the epistemic significance associated with the (counterfactual) results that a scientist will accrue from pursuing the approach.

In their simulation, Weisberg and Muldoon use a three dimensional representation: a two-dimensional configuration space for approaches, and a third dimension, represented as height, for significance. In all runs of the simulation the same landscape is used, with the following parameters:

size 101x101 discrete toroidal (left and top edges "loop back" to right and bottom edges) grid.

baseline significance at baseline level is 0.

peaks the landscape contains two Gaussian peaks.

This choice of landscape structure leaves a significant portion of the landscape at baseline (0 significance) level. The use of a Gaussian structure for the peaks creates a very smooth rise in significance where significance is non-zero. Weisberg and Muldoon seem to be concerned with a situation where only one or two approaches (or clusters of similar approaches) to a topic can yield any significant results, while a large number of alternative approaches will yield no significant results; they do not consider the alternative of a field where almost any approach will yield *some* significant result, but there are specific approaches within this multitude which are locally better than their alternatives (i.e. there are many good approaches to learning about this topic, but if you choose any of these approaches there is a best method for doing so). The ruggedness of the terrain, and its potential influence on the epistemic landscape model, are discussed later in the chapter.

5.1.3 Scientist agents

Weisberg and Muldoon's epistemic landscape is populated with simulated agents. These represent individual scientists, such that an agent at a certain coordinate of the landscape in a certain simulation step represents a scientist pursuing a certain approach at a certain time. The agents are therefore identified by their location, or the approach they are pursuing. In addition, each agent is assigned a "character", which corresponds to the algorithm the agent executes each step to update her position on the landscape (overall there are three options for "character" algorithm, with no variation between agents who have the same "character"). At each run of the simulation, the agents are seeded in random locations of zero significance on the landscape, and then in every step each agent's position is updated using that agent's characteristic algorithm.

5.1.4 Simulation and measurement

Weisberg and Muldoon's simulation is conducted by generating the epistemic landscape mentioned above, seeding it with a population of agents with certain characters, either pure or mixed, and running the simulation such that in every step each agent carries out their movement algorithm. The performance of different population sizes and characters was then compared using several measurement parameters:

Are maxima reached The first, and most basic, simulation tested whether populations manage to find either a) the global maximum, or b) all local maxima.

Time to reach maxima Next, Weisberg and Muldoon tested the time required to reach all local maxima.

Epistemic progress A measure of a population's ability to find significant approaches, epistemic progress is defined as the percentage of non-zero-significance patches visited at any point during a fixed-length simulation run. Populations are considered more successful if, for a set number of steps, they reach higher epistemic progress.

Total progress A measure of the thoroughness of investigation, total progress is defined as the percentage of patches, regardless of their significance, visited at any point during a fixed-length simulation run.

Perhaps surprisingly, Weisberg and Muldoon do not measure the *height* of patches visited. Thus, even though they assign numerical values of significance, these are only used to determine the behaviour of agents, and are not integrated into a measure of the success of the population. This decision is not motivated explicitly in their work; we could speculate that this choice represents a lack of confidence in the "significance" parameter as tracking something real, and so they prefer to evaluate the success of the population of investigators using parameters that are more readily associated with reality, such as time, or volume of scientific output.

As discussed in Chapter 2, Weisberg and Muldoon's simulation is not appropriate for modelling science funding decisions. Briefly, it fails to represent the change in significance over time that results from the agents' movement on the landscape. Nonetheless, much of the machinery used by Weisberg and Muldoon is very promising, and has been co-opted into the simulation presented in this chapter, as discussed below.

5.2 Simulation description

The simulation presented in this chapter is aimed at the exploration of the influence of different funding mechanisms on the epistemic progress, defined as the accumulation of results of high fitness, achieved by a population of investigators. Following the conceptual and historical analysis of the previous chapters, it is suggested that the simulation should include processes by which the fitness of projects can change over time, as a response to research activity. To pursue this aim, a simulation is presented that builds on Weisberg and Muldoon's simulation presented above, while drawing on elements from population biology simulations to extend the simulation in the directions of funding and dynamic fitness.

The simulation simulates a population of scientists exploring a topic of scientific interest. They are all funded by a central funding body to pursue projects of varying duration (represented

by the number of simulation steps required to finish a project, each step roughly corresponding to one year). Each scientist pursues the project for which they were funded for the duration of the grant period, at the end of which they obtain the results of the project, which are allocated in advance by the simulation from a "god's-eye" perspective. These results are then incorporated into the scientific corpus associated with the field, which leads to an increase of fitness for that corpus. Funding mechanisms are compared by their ability to generate this increase in fitness.

The "god's-eye" perspective mentioned above can be conceptualised using the definition of epistemic fitness given in §3.2.3. Since fitness measures the extent to which the causal consequences of a research project's results will help the society progress in the direction defined by the model-user's notion of well-being, a "god's-eye" perspective would be held by an evaluator who has access to these two sets of information, i.e. the totality of the causal consequences of the results and a specific way of assigning to these consequences an evaluation of contribution to well-being. A reasonable approximation of this "god's-eye" perspective would be the perspective of a historian researching the impact of a scientific discovery on a society many decades after its publication, assuming the scholar has infinite time to gather information, and that the society has maintained sufficient documentation of events and opinions regarding these events.

For the sake of simplicity, the activity of the agents is divided into two parts. In the first part, which takes place throughout the duration of the grant, each scientist is "investigating" a single project, accumulating data and testing hypotheses, but they do not share their findings nor explore similar projects. It is as if they have gone into the wilderness, or locked themselves in the lab, for as long as money would last, in the pursuit of new knowledge on a particular narrow topic. In the second stage, which in the simulation takes place instantaneously, the results accumulated throughout the duration of the investigation are "submitted" to the corpus associated with the research field, causing an instantaneous recognition of the importance of the results to well-being. This is represented in the simulation as an increase in the fitness of the corpus. This is akin to the scientist emerging from the edge of the forest, or coming out of the darkness of the lab, holding a detailed report and documented evidence to support it. While this picture is clearly a gross simplification of reality, it is necessary in order to focus on the main aim of the simulation, which is to explore how a community containing *multiple* scientists and *multiple* paths into the wilderness should be optimally organised.

To simulate the effects of funding, the population of scientists goes through a process akin to selection, or more accurately, akin to selective breeding. In every step of the simulation, the scientists whose grants have run out are placed in a pool of candidates along with new entrants to the field, and the simulated funding mechanism selects from this pool of candidates those who will receive funding and carry out research projects. The different funding mechanisms differ in the way they select individuals, the details of which are described in §5.3.3. Overall, the simulation compares funding based on:

Actual potential Actual potential, which can only be known from a god's-eye perspective, is the fitness contribution associated with the results of a project *were it completed today*. In the absence of fitness dynamics, actual potential is simply the fitness contribution of the project. However, in the presence of fitness dynamics, there is a non-zero chance that the fitness of the project will change between the initiation of the project (at the point of funding) and its completion (at the point of contributing the result to the relevant corpus). This means that in the presence of fitness dynamics, actual potential (evaluated at the

point of funding) might diverge from the eventual fitness contribution of the project.

Estimated potential Estimated potential is the scientific community's prediction (assumed to be single-valued) of the fitness contribution of a proposed project. This prediction is taken to rely on the known fitness contributions of past projects which bear some similarity to the proposed project. Thus, estimated potential depends on the history of research projects in the field. In representing decisions based on the research community's prediction, this selection method is akin to peer-review, as described by Polanyi (see §1.5).

Lottery Under a lottery, all candidates have equal chances of being funded, as proposed by Gillies (2014) (see §1.6). There are several reasons motivating the inclusion of the lottery option in simulations: first, it seems intuitive that we would want any funding mechanism implemented to outperform a lottery, and so we need to simulate the performance of a lottery mechanism for comparison; second, the lottery presents a form of scepticism, which follows from the discussion in the previous chapter; finally, the lottery is an extreme "anti-conservative" option, which contrasts well with the "past funding" option. For all of these reasons, it is expected that the inclusion of a lottery mechanism in the simulation will prove informative.

Past funding Under this mechanism, funding is allocated to those scientists who already received funding in the past, and only to them. As the simulation currently treats all scientists to be alike, this mechanism cannot be taken to mean the selection of the most "intrinsically able" scientists. Rather, this mechanism is included as a "most conservative" option, which is effectively equivalent to not admitting any new researchers to the field beyond the field's original investigators.

The essence of the simulation is in the comparison of the performance of these methods in generating results of high fitness over time under various conditions.

To simulate dynamic fitness, the fitness contributions of different projects are allowed to change over time as a response to scientists' actions. Specifically, three processes of fitness dynamics are simulated, the details of which are given in §5.3.4. Two processes involve the reduction of fitness differential that follows a successful project or breakthrough, which reflects the one-off nature of discovery: there is little interest in discovering a novel phenomenon a second time, for it is no longer novel. The third process involves the increase in fitness of a group of projects which represent a new avenue of research being opened by a successful and significant discovery. The simulation shows that the presence of these processes has a significant effect on the relative performance of different funding strategies.

The next section presents a detailed description of the simulation components and parameters. The section following provides a presentation of simulation results and discussion thereof.

5.3 Simulation details

5.3.1 Simulating the epistemic landscape

§3.5 presented a revised version of Weisberg and Muldoon's epistemic landscape model (Weisberg and Muldoon, 2009). That chapter argued that the epistemic landscape has much in common with the fitness landscape model in biology. Thus, in simulation, the starting point is to make a

model that resembles both Weisberg and Muldoon's simulation and the models which have been used in evolutionary biology to simulate evolutionary processes on fitness landscapes.

The basic structure of the landscape simulation follows Weisberg and Muldoon's, of a two-dimensional configuration space, charted with two coordinates x and y, with an associated scalar field represented in a third dimension as height along the z axis. The scalar field, which in Weisberg and Muldoon's model represents "significance", represents "epistemic fitness differential", using the notion of epistemic fitness introduced in §3.2.3.

As mentioned by Weisberg and Muldoon, and expanded in Chapter 3, the choice of a two-dimensional configuration space is very likely to be a simplification, with the actual number of dimensions for the landscape (if countable, discrete dimensions can be assigned to it at all) being much greater than two. However, for ease of calculation and visualisation, a two dimensional configuration landscape offers great advantages over a higher-dimensional one. This does mean, however, that when dynamics are explored and results are interpreted, we need to ask ourselves how these would behave in a higher-dimensional corollary. The high-dimensional nature of the target domain, and the wish to maintain similarity to Weisberg and Muldoon's model, as well as biological models of fitness landscapes, motivated the choice of a two-dimensional configuration space over the simpler one-dimensional alternative.

In each run of the simulation, the landscape is generated anew in the following process:

1. Initialise a flat surface of the required dimensions.
2. Choose a random location on the surface.
3. Pick random values for relative height, width along x, and width along y.
4. Add to the landscape a hill at the location chosen in step 2 by using a bivariate Gaussian distribution with the parameters picked in step 3.
5. Repeat steps 2-4 until the specified number of hills is reached.
6. Scale up linearly the height of the landscape according to the specified maximum height.

Some examples of landscapes are shown in Fig. 5.1 and Fig. 5.2.

As the process uses random variables in several steps, each run generates a slightly different landscape. Random variables are useful for the exploration of strategies (discussed later) on a variety of landscapes, as the actual shape of epistemic landscapes varies. In general in the simulation, random variables are used to fill-in parameters whose existence is essential for the simulation, but where a) the specific values they take can vary across a range of valid model targets, and/or b) there is no compelling empirical evidence to choose a particular value. This requires, however, several runs of the simulation for each strategy, to average out unwanted effects that might result from random variation.

The key parameters that determine the shape of the landscape are its size, maximum height, and characteristics of hills. For simplicity all landscapes are square, so size is given as the length of one side, e.g. size of 100 corresponds to a landscape with coordinates ranging from (0,0) to (100,100). For ease of visualisation, the maximum height is set as twice the size.[1] The hills are characterised by their height and their spread along x and y.

[1] The specific maximum height is not expected to have a significant effect on the model.

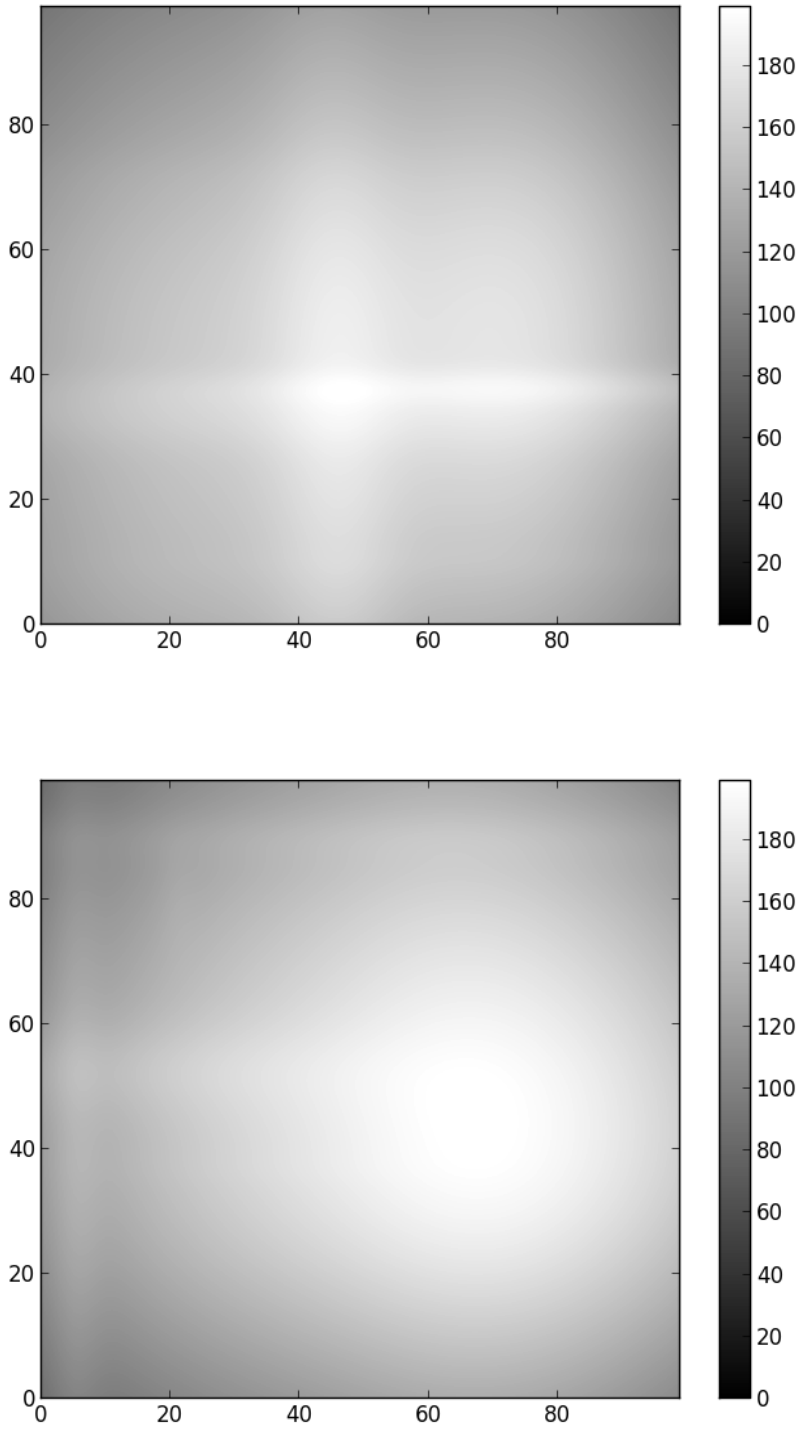


Figure 5.1: Two landscape simulations, with colour bar. Each simulation has size of 100x100, with 50 hills, and maximum height of 200. The dominant feature is a broad "hill".

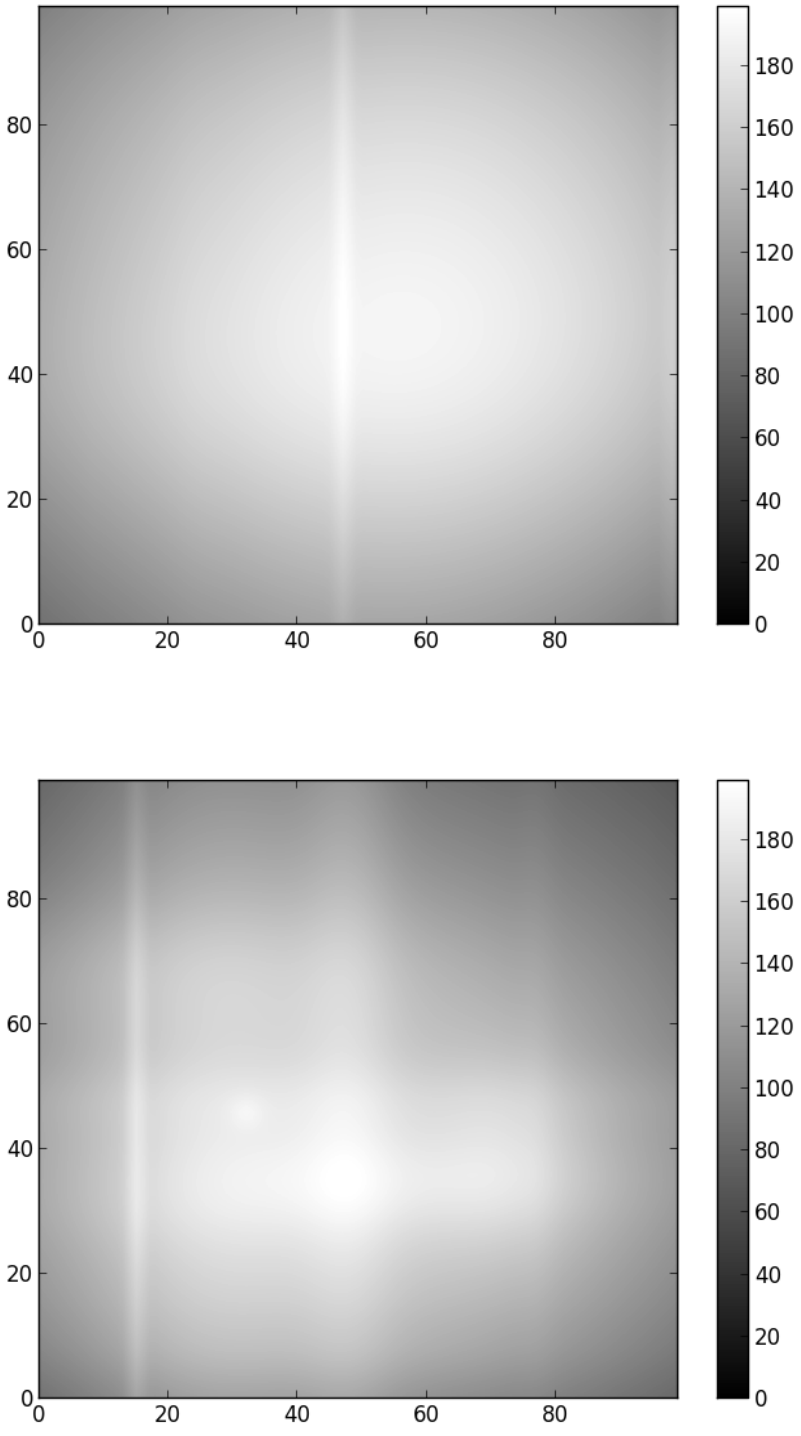


Figure 5.2: Two landscape simulations, with colour bar. Each simulation has size of 100x100, with 50 hills, and maximum height of 200. The dominant feature is a set of "ridges".

Ruggedness of the epistemic landscape

In keeping with Weisberg and Muldoon's simulation, I simulate an epistemic landscape using Gaussian hills. The number of hills mentioned above refer to the process of generating the landscape, not the final product: as can be seen in Figs. 5.1 and 5.2, the use of 50 hills in the generation process yields only two to four peaks on a rather smooth landscape. This is an expected statistical result, where the sum of independent Gaussians tends itself to be a Gaussian. Unlike Weisberg and Muldoon's landscape, in my method of generating the landscape I allow for variability in the structure of the landscape (e.g. hills versus ridges), and also cover most of the landscape (statistically) with positive fitness. These features follow from the generation of the epistemic landscape as a derivative of the information landscape, under the assumption that the information landscape itself is relatively smooth (see Chapter 3).

In the context of thinking about fitness landscapes, Kauffman (1993) has shown, using the NK model, that the degree of ruggedness (roughly equivalent to the number of local maxima) has a direct effect on the course of evolution. As evolving agents are forced, by selection, into local hill-climbing, the presence of many local maxima greatly increases the chances for a population to get "stuck", i.e. for it to reach a local maximum and from that point onwards stay at that maximum, never reaching higher maxima that may exist on the landscape. Similar arguments could be advanced in the case of the epistemic landscape. Nonetheless, the strength of these arguments relies on the fact that the fate of any single organism under natural selection only depends on the specific fitness at the current location of that organism, and is not influenced by the fitness of other organisms on the landscape, or on the fitness of ancestors. As described below, only one of the funding mechanisms is simulated as "pure" hill climbing, which only takes into account the local fitness of the project each agent is pursuing. The other selection mechanisms simulated take into account non-local aspects of the landscape, and are thus less susceptible to getting agents "stuck" (more details below). Furthermore, the addition of landscape dynamics makes the problem of local maxima less severe, as a "stuck" agent may become unstuck due to changes of topology in its local area (again, more on this below). Thus, issues arising from ruggedness seem to be of less importance in the simulations presented in this chapter than in other fitness landscape simulations.

5.3.2 Simulating agents

The agents in the model represent scientists who are investigating the topic represented by the epistemic landscape. Each agent represents an independent researcher or group, and is characterised by its location on the landscape, representing the project they are currently pursuing, and a countdown counter, representing the time remaining until the current project is finished. Agents are simulated as greedy local maximisers.[2] Agents follow the following strategy every simulation step:

- Reduce countdown by 1.
- If countdown is not zero: remain in same location.

[2]Here and later in the chapter, "greedy" is used as a technical term to describe (an agent following) a greedy algorithm, not to be confused with the colloquial use of the term. No assumption or judgement is made about the moral character of the agents or their motivations. For an introduction to greedy algorithms see Cormen (2001).

- If countdown is zero: contribute to the global fitness the fitness of the current location, and attempt to move to the highest local neighbour.

The *local neighbourhood* of an agent is defined as the 3x3 square centred on their current position. The attempt to move to the highest neighbour depends on the selection (funding) mechanism, as discussed below. The accumulation of *global fitness*, which is the sum of all individual contributions to fitness when an agent's countdown reaches zero, is stored as a global variable of the simulation and used to compare strategies, as discussed below.

In the beginning of the simulation, a specified number of agents are seeded in random locations on the landscape, with randomly generated countdowns selected from a specified range of values. An example of an initial seeding of agents can be seen in Fig. 5.3.

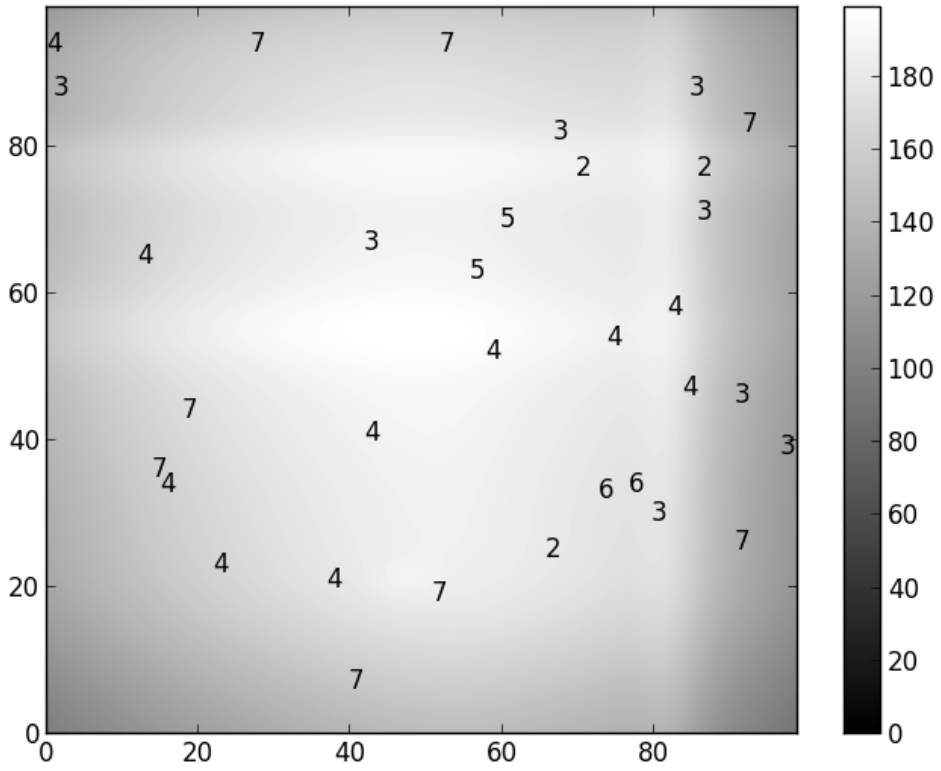


Figure 5.3: Landscape simulation with initial seeding of agents, with colour bar. Each number on the landscape represents an agent at its location, with the value of the number representing the agent's countdown.

In addition to their contribution to fitness, agents also contribute to the *visibility* of the landscape. The visibility of a project represents whether the scientific community, and especially funding bodies, can estimate the fitness contribution, represented in the model as the height, of the project. Initially, the entire landscape is invisible, representing full uncertainty about the fitness of the possible projects in the topic. Upon initial seeding of agents, each agent contributes vision of their local neighbourhood, as defined above, to the total vision. As the agents move, they add vision of their new local neighbourhood. In the simulation, vision of a certain project is never lost, even when its fitness changes, as described below. Thus, as the simulation progresses, vision of the landscape increases monotonously, covering an increasingly larger portion of the

landscape as the agents move about. Visibility is used in the *best_visible* funding mechanism described below.

The key parameters that govern the simulation of agents are the number of agents relative to the size of the landscape, the average length of a project, represented by the countdown value, and the spread of possible project lengths. The number of agents, relative to the size of the landscape, affects the model in more than one way: significantly, it determines how long an agent can continue in a research programme (series of projects) before encountering the path of another agent, i.e. the agent's *mean free path*. In addition, because of the fitness dynamics described below, the number of agents influences the probability that the fitness of projects in an agent's local neighbourhood will change in a given time delta, with the more agents, the greater the chance of a dynamic effect.

In the absence of selection and dynamics, the course of the simulation is easy to describe: agents begin in random locations on a random landscape, and as the simulation progresses the agents finish projects and climb local hills, until, after an amount of time which depends on the size of the landscape, the number and size of peaks, and the length of projects, all agents trace a path to their local maxima and stay there. Since agents increase their local fitness during the climb, the rate of fitness gathering increases initially, until all agents reach their local maxima, at which point the rate of fitness acquisition stabilises and global fitness increases linearly with a fixed rate. The global fitness as a function of time for such a system is charted in Fig. 5.4.

5.3.3 Simulating funding strategies

One of the two main aims of the simulation is to simulate the effect of funding mechanisms on the population and distribution of investigators. Since the aim is to simulate current funding practices (albeit in a highly idealised manner), and since current funding practices operate in passive mode (see Chapter 1), the guiding principle of the simulation is that a funding mechanism is akin to a selection processes in an evolutionary simulation: at each step of the simulation, the actual population of agents is a subset of the candidate or potential population, where inclusion in the actual population follows a certain selection mechanism. This guiding principle provides further motivation for using a simulation which is analogous to the simulations used in population biology.

The inclusion of a funding mechanism is simulated in the following manner:
Every step:

- Place all agents with zero countdown in a pool of "old candidates".
- Generate a set of new candidate agents, in a process identical to the seeding of agents in the beginning of the simulation.
- Select from the joint pool of (old candidates + new candidates) a subset according to the selection mechanism specified by the funding method.
- Only selected agents are placed on the landscape and take part in the remainder of the simulation, the rest are ignored.

The simulation can represent four different funding mechanisms:

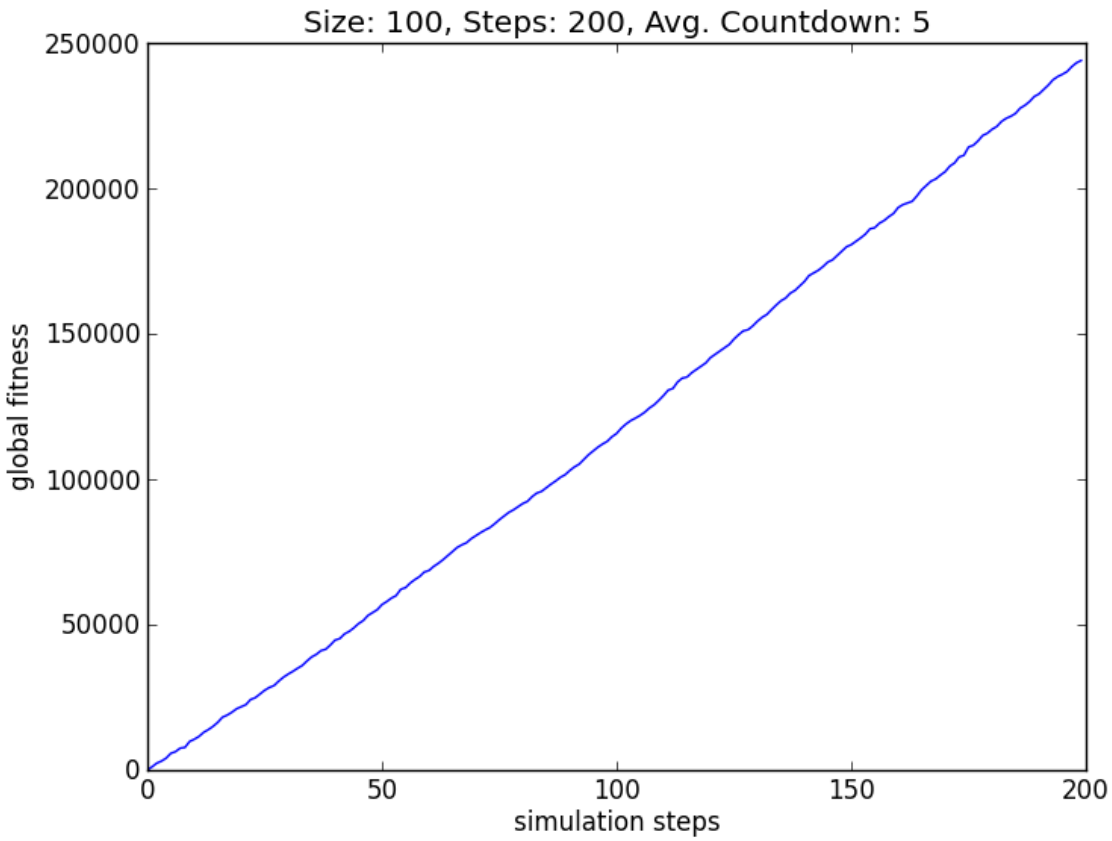


Figure 5.4: Global fitness accumulated as a function of time, represented by simulation steps, for a simulation with no selection mechanism and no fitness dynamics. The graph represents local hill climbing, with the rate of fitness accumulation increasing initially and then stabilising as each agent reaches their local maximum. The simulation was run for 200 steps on a landscape of size 100x100 with maximum height of 200, with 31 agents with countdowns ranging from 2 to 7.

best selects the candidates which are located at the highest points, regardless of the visibility of their locations. This simulates a mechanism which selects the most promising projects, *ex ante*, from a god's eye perspective. Since actual visibility is limited, this overly optimistic mechanism does not represent the strategies of real funding bodies, though it may serve as an ideal benchmark against which realistic funding mechanisms may be measured.

best_visible filters out candidates which are located at invisible locations, i.e. candidates who propose to work on projects which are too different from present or past projects, and then selects the candidates in the highest locations from the remainder. This strategy, with the visibility constraint, is much closer to a representation of the funding strategies of current funding bodies. Note that even this version is somewhat epistemically optimistic, as it assumes the funding body has successfully gathered all the available information from all the different agents, both past and present, which is the epistemic requirement for accessing the global visibility map (the visibility map is described in §5.3.2).

lotto selects candidates at random from the candidate pool, disregarding the visibility and height of their locations.

oldboys represents no selection: old candidates continue, no new candidates are generated. This would correspond, for example, to a field explored only be tenured academics, each continuing on their own research programme indefinitely (though possibly indirectly interacting with the other tenured academics via fitness dynamics), with no new entrants to the field.

The key parameters for all funding mechanisms are the size of the candidate pool and the size of the selection pool. The size of the candidate pool, which in turn depends on the size of the new candidate pool (as the size of the old candidate pool emerges from the simulation), has been chosen in the simulations such that the total candidate pool equals in size to the initial number of agents, except in the *oldboys* option in which it equals the size of the old candidate pool. This means the success rate for an individual candidate in each funding round changes somewhat, around a mean which is equal to 1/(average countdown). Since in most simulations the average countdown used is 5, this yields a success rate of 20% which is close to the real value in many contemporary funding schemes (NIH, 2014). The size of the selection pool is set to equal the size of the old candidates pool, representing a fixed size for the population of investigators. This too corresponds well to many contemporary funding schemes, though it was not so in the past.

An important simplifying assumption is that the funding mechanisms do not take into account the positions of existing agents on the landscape, except indirectly when considering their vision. This is done for simplicity, and a more complex simulation may consider a selection mechanism which explicitly favours either diversity or agglomeration. However, to some extent the position of existing agents affects the potential for funding of new candidates by influencing the fitness landscape itself, as discussed below.

5.3.4 Simulating fitness dynamics

In addition to simulating funding mechanisms, the other main aim of the model is to simulate the changes in fitness that occur as the result of investigations. The motivation for exploring fitness dynamics was discussed at length in the previous chapter. To summarise, historical evidence strongly suggests that the fitnesses of projects can and have changed, sometimes significantly,

over short periods of time. Since funding strategies are evaluated based on their ability to accumulate global fitness, and the accumulation of fitness depends on the specific fitness values of projects at specific times, which may change, it is important to look at fitness dynamics when evaluating funding strategies.

In the presence of fitness dynamics, it is problematic to speak of "the accumulation of fitness", because the fitness contributions assigned to past projects may change as a result of new research. What the simulation tracks in these cases is the quantity of fitness contribution estimated at the time when the project is completed and the results are clear. While this value will not be fixed for all future times, this juncture does represent a significant drop in the uncertainty associated with the fitness gain assigned to the project. The simulation compares funding strategies based on their ability to predict what the estimation of fitness will be at the project's completion, a prediction made at the beginning of the grant, i.e. several years into the future. This is a major step away from the earlier version of the sceptical argument, which argues against funding bodies' ability to maximise *eventual* contributions to well-being. However, even this weaker version succeeds in showing that maximisation strategies are sometimes undermined by fitness dynamics, even if we only consider fitness changes *over the duration of a single grant period.*

In the current simulation, which is limited to a single landscape (unlike the many landscape model discussed in Chapter 3), only inter-topic processes of fitness dynamics are simulated (see §4.3). Specifically, at present only three processes are simulated:

Winner takes it all As was made explicit by Strevens (2003), the utility gain of discovery is a one-off event: the first (recognised) discovery of X may greatly contribute to the collective utility, but there is little or no contribution from further discoveries of X. In the simulation, this is represented by setting the fitness of a location to zero whenever an agent at that location has finished their project (countdown reaches zero) and gained the fitness from that location. This effect is triggered whenever any countdown reaches zero, which makes it quite common, but it has a very localised effect, only affecting the fitness of a single project.

Reduced novelty When an agent makes a significant discovery, simulated by finishing a project with associated fitness above a certain threshold, the novelty of nearby projects is reduced, which in the model is simulated by a reduction of fitness in a local area around the discovery.

New avenues When an agent makes a significant discovery, it opens up the possibility of new avenues of research, simulated in the model by the appearance of a new randomly-shaped hill in a random location on the landscape.

The key parameters that characterise all dynamic processes are their trigger, their magnitude, and their location. In the current simulation there are two types of triggers: a process is either triggered whenever a project finishes, or only when a project above a certain fitness threshold finishes. Effects are simulated as bivariate Gaussians, so their magnitudes are determined by the height of the peak and the spread along x and y, where these values are selected at random within a given range. There are two options for the location of the effect: it is either localised at the trigger, or else it is placed at a random location on the landscape, with each position having an equal probability.

The results of the simulation are presented below.

5.4 Results and discussion

Here I present the results of simulations of five setups of interest, exploring the effect of fitness dynamics on the success of different funding mechanisms.

All simulations show a comparison of the four funding mechanisms, as a plot of accumulated fitness over time. The lines are averages of five runs. In all simulations the range of countdowns was 2 to 7. The number of individuals was set to equal (size of landscape)$^{3/4}$. In simulations with dynamic processes, the trigger for fitness-dependant processes was 0.7 of the global maximum.

The first four results relate to the four quadrants of Table 5.1 and depict simulations of 50 steps. The fifth result, shown in Fig. 5.12 simulates a long evolution (200 steps) on a medium sized dynamic landscape.

Table 5.1: Simulation setups and associated figures

	static landscape	dynamic landscape
small landscape	Fig. 5.8	Fig. 5.10
large landscape	Fig. 5.9	Fig. 5.11

To enable a richer understanding of the simulation, a few scenarios have been visualised. These are presented in Figs. 5.5, 5.6, and 5.7.

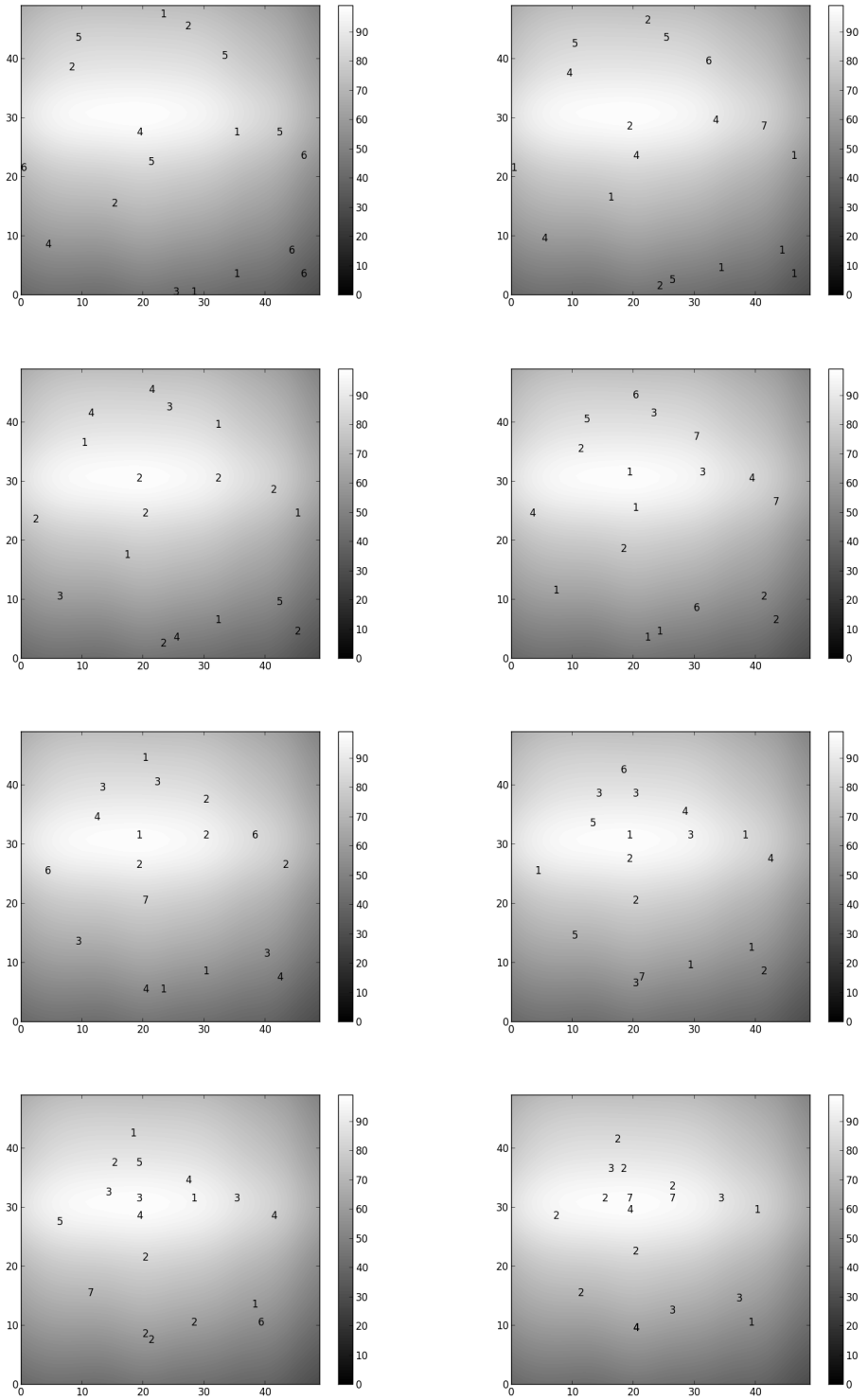


Figure 5.5: Simulation of *oldboys* funding mechanism on a static 50x50 landscape. The simulation progresses from top-left to bottom-right, with consecutive panels showing snapshots of the simulation at five step intervals.

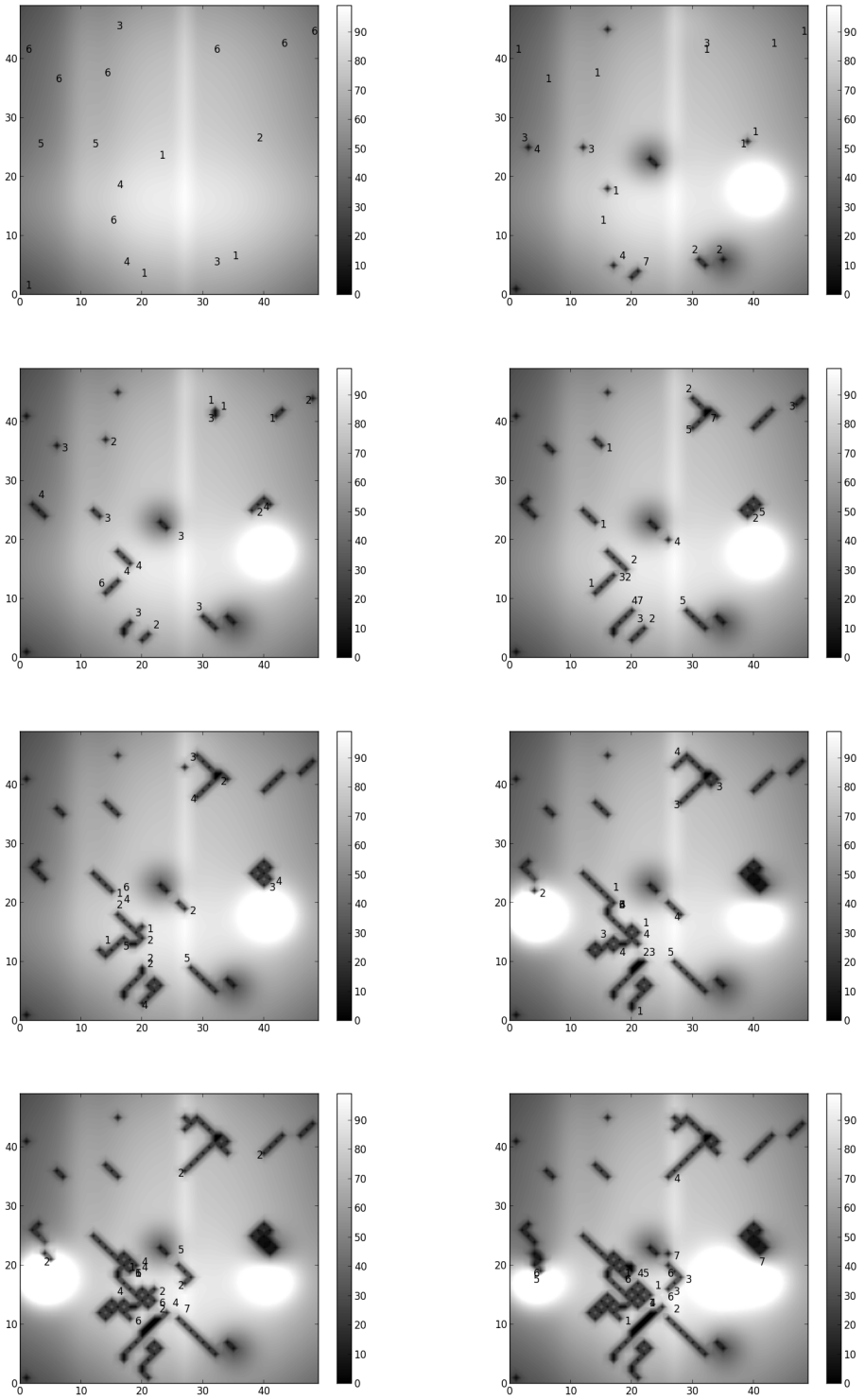


Figure 5.6: Simulation of *best_visible* funding mechanism on a dynamic 50x50 landscape.

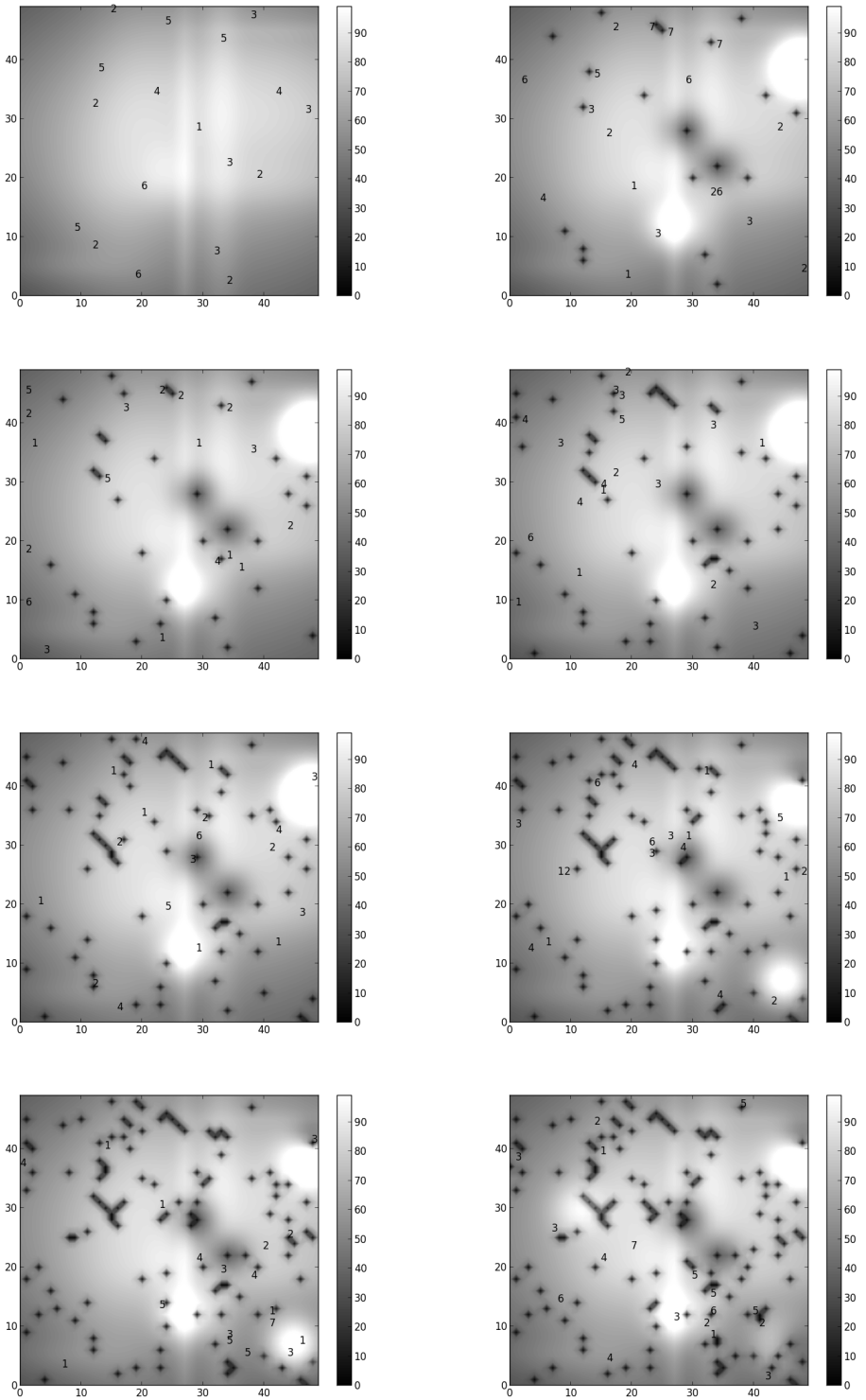


Figure 5.7: Simulation of *lotto* funding mechanism on a dynamic 50x50 landscape.

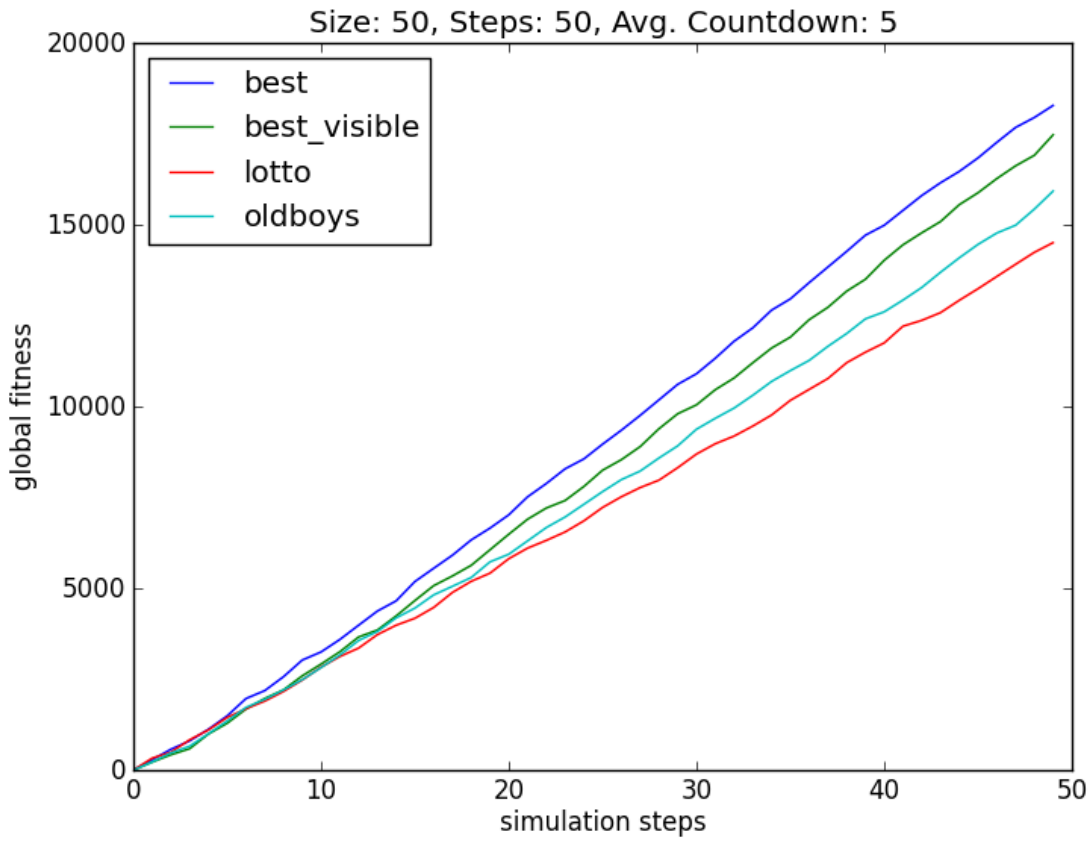


Figure 5.8: Comparison of different funding mechanisms on a small (50x50) landscape with no dynamic processes.

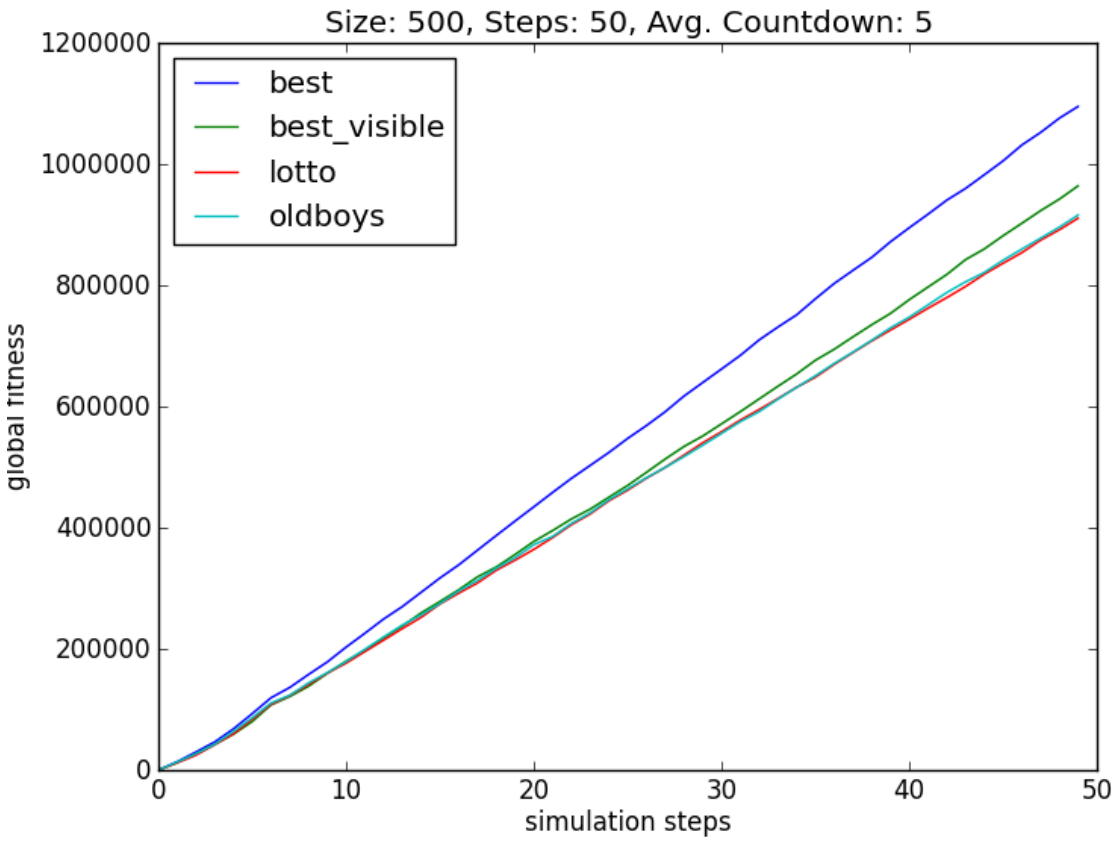


Figure 5.9: Comparison of different funding mechanisms on a large (500x500) landscape with no dynamic processes.

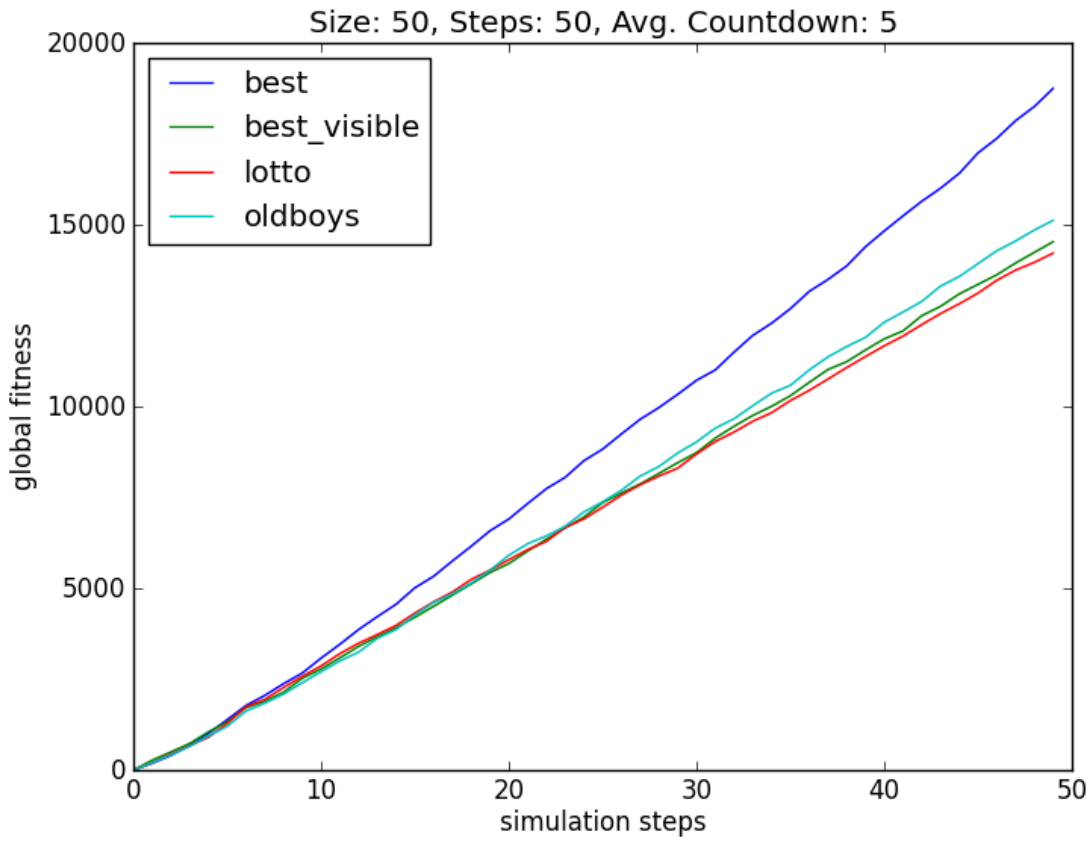


Figure 5.10: Comparison of different funding mechanisms on a small (50x50) landscape with dynamic processes.

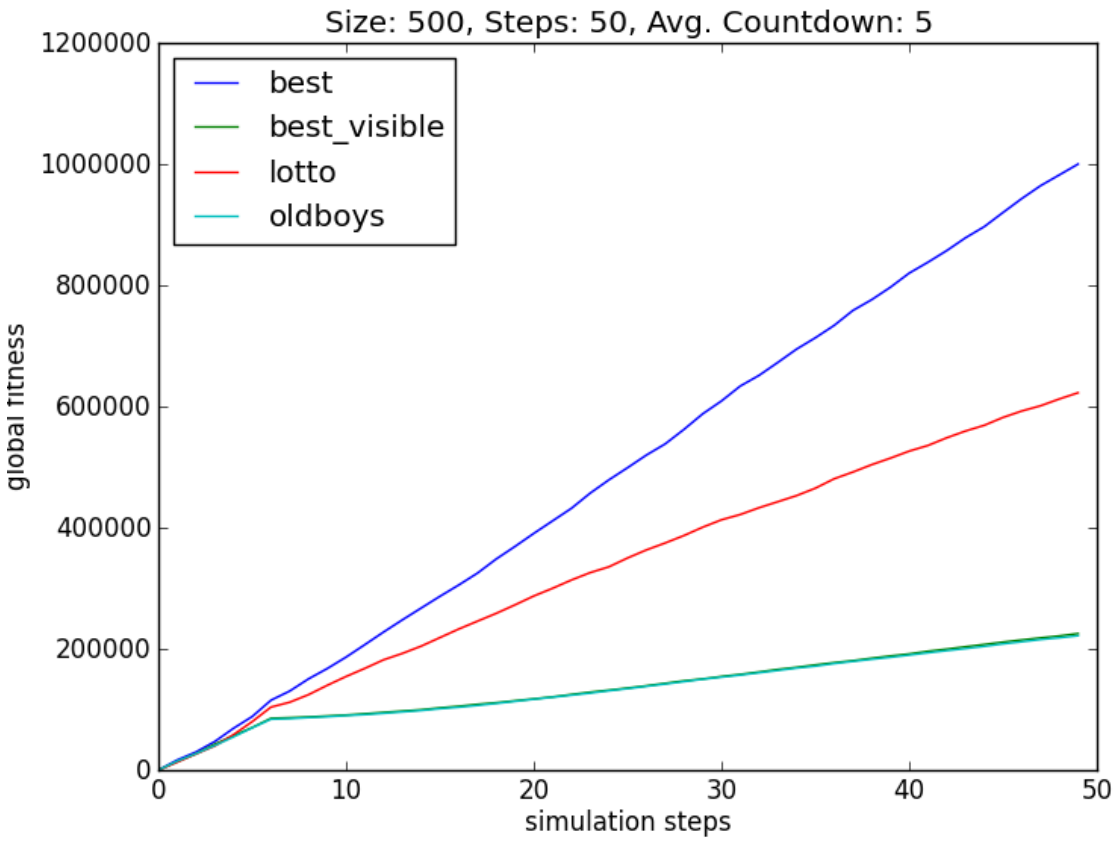


Figure 5.11: Comparison of different funding mechanisms on a large (500x500) landscape with dynamic processes. Note that *best_visible* and *oldboys* are almost entirely overlapping.

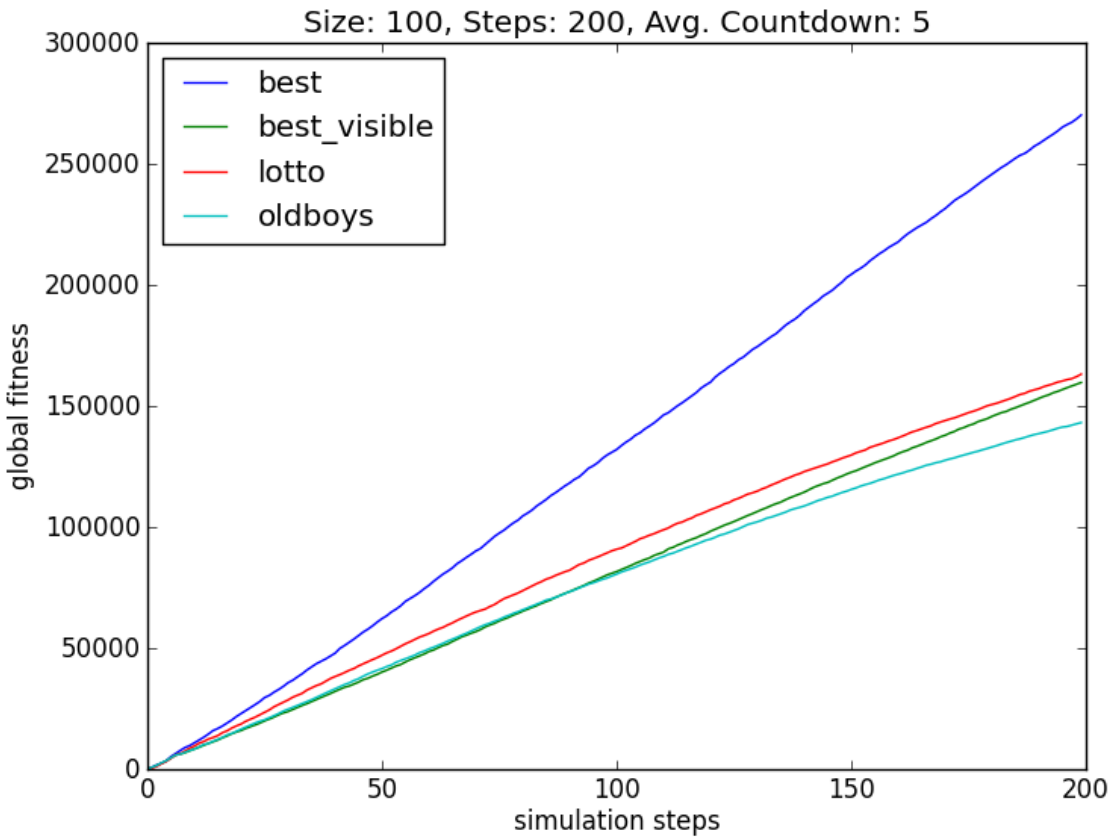


Figure 5.12: Comparison of different funding mechanisms on a medium sized landscape (100x100) with dynamic processes over a long period of time (200 steps).

As is clear from all simulations, the *best* funding mechanism is indeed best at accumulating fitness over time, though with various lead margins over the second best strategy. On static landscapes, *best* has access to global maxima, unlike *best_visible* and *oldboys*, and can preferentially select agents which work on projects near or at global maxima, unlike *lotto*. On dynamic landscapes, *best* is in the best position to locate new avenues for research, wherever they show up. However, as mentioned above, the *best* funding strategy is not realisable, as it requires a god's eye view of the epistemic landscape.

On the static landscape, *best_visible* outperforms *lotto* and *oldboys*. This is expected; on a static landscape *oldboys* performs a greedy local maximisation, and *best_visible* provides a shortcut for individuals to "jump" to the highest visible point, shortening the duration of the climb and reaching the local maxima faster. On static landscapes *lotto* does poorly, as agents do not maintain their positions on maxima even if reached, but are rather replaced with randomly placed agents who are likely to spawn on suboptimal positions.

On the dynamic landscape the results are different. On the small landscape the three strategies, *best_visible*, *oldboys*, and *lotto* perform roughly similarly, with the dynamic effects offsetting to an extent the disadvantages of *lotto*. However, the significant effect is seen on the large dynamic landscape, where *lotto* greatly outperforms *best_visible* and *oldboys*. This is because new avenues on a large landscape are likely to spawn outside the visibility of the agents, where *lotto* can access them but the other two strategies cannot. In the smaller landscape this effect is not apparent, as the relative visibility is larger, and therefore the chance of a new avenue appearing within the visible area is larger. This result is discussed in detail in the next chapter.

Finally, the long duration simulation shows the slowing rate of fitness accumulation over time, given fitness dynamics. This result agrees with the intuitive assessment made by Peirce (1879/1967), that inquiry in a given area will display diminishing returns.

The surprising success of *lotto* in outperforming *best_visible* suggests a further simulation, in which the selection of candidates carries features from both these methods. This new funding mechanism, labelled *triage*, was simulated by filling half the open positions at random from outside the vision range (similar to the *lotto* mechanism), and half by selecting the best candidates from within the vision range (similar to *best_visible*). An example of the progression of the simulation under *triage* is shown in Fig. 5.13. This new mechanism was then compared to the other funding mechanisms, on both small and large dynamic landscapes. The results are shown in Fig. 5.14 for a small landscape, and in Fig. 5.15 for a large landscape. The *triage* option is discussed in detail in the next chapter.

The following sections deal with possible criticisms of the simulation, and replies to these criticisms.

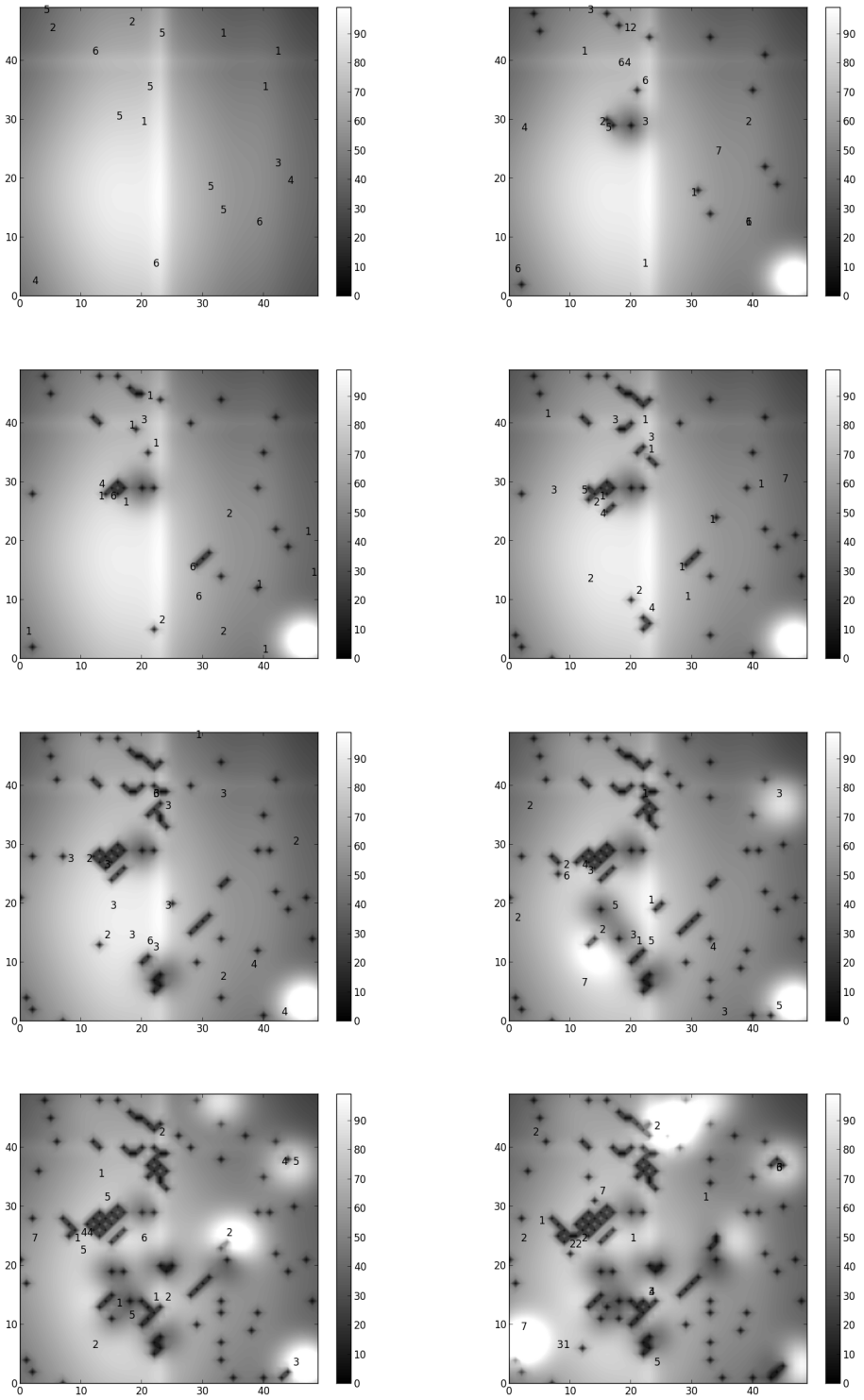


Figure 5.13: Simulation of *triage* funding mechanism on a dynamic 50x50 landscape.

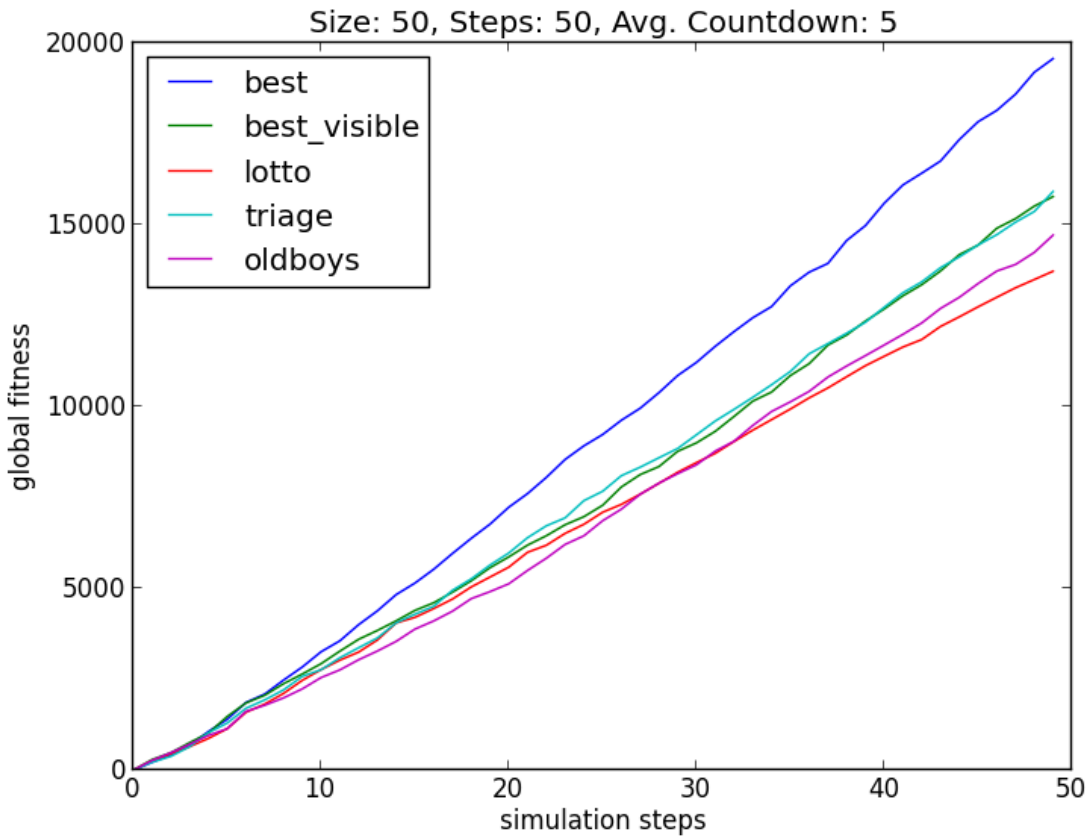


Figure 5.14: Comparison of different funding mechanisms, including *triage*, on a small (50x50) landscape with dynamic processes.

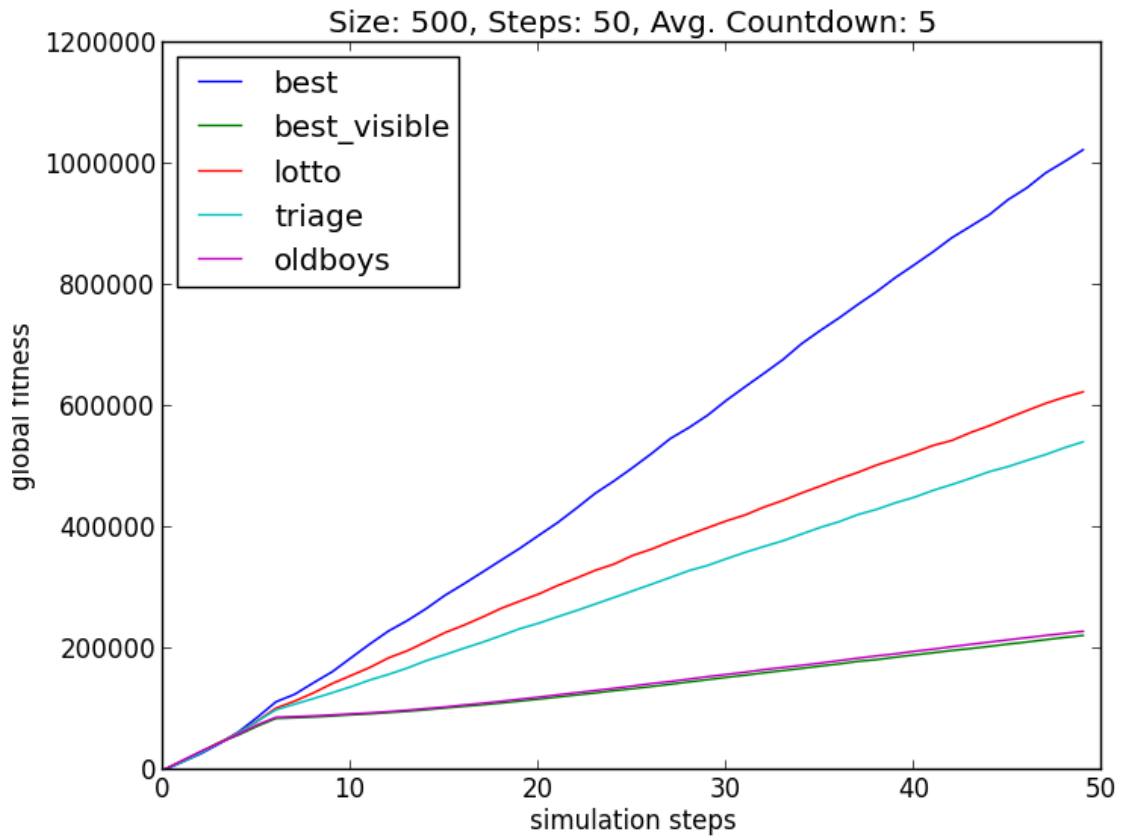


Figure 5.15: Comparison of different funding mechanisms, including *triage*, on a large (500x500) landscape with dynamic processes.

5.5 Simulating in a data-poor domain

Since the simulation does not rely on empirical data, what epistemic status does it have?

The simulations presented in this chapter are not driven by empirical data, but by idealised representations of "reasonable assumptions" about the target domain. Therefore, it may be useful to think of the simulation as a formalised thought experiment. Both the thought experiment and this idealised simulation operate by making a complex system *concrete* (in the sense of specific, not in the sense of actual). However, unlike the thought experiment, which concretises by loading a hypothetical anecdote with what are taken to be exemplary characteristics, the simulation concretises by assigning numerical parameters to what are taken to be key processes.

Take for example the famous switch-trolley thought experiment (Foot, 2002). In this thought experiment, an out-of-control trolley is heading towards five persons chained to the tracks. A lever-operated switch will alter the course of the trolley so it will instead go on a rail to which only one person is chained. The question: given influence only on the lever, will you pull it and switch the direction of the trolley? This thought experiment is meant to concretise general ethical concerns to do with expected consequences, responsibility and agency, and to hone in on specific ethical criteria.

In the design of both the thought experiment and the simulation, the usefulness of the concretisation relies on our judgements of what is reasonable and what is important. When a philosopher poses a thought experiment to explore ethical decisions, for example, her judgement is called upon twice: first, she needs to decide on a specific context, e.g. a person standing by a switch near the tracks rather than, say, a military commander in an operations room deciding on whether to launch the missiles; second, she needs to decide which characteristics will be considered relevant in the chosen context, e.g. to include the number of people chained to the tracks but exclude details about the people's lives, characters or past actions. In the design of a simulation, the designer's judgement is also called upon twice: in the choice of relevant processes to include in the simulation, and in their numerical parametrisation.

Due to their reliance on largely untested beliefs about what is relevant and what is reasonable, both the thought experiment and the simulation are not predictive. Nonetheless, in the best scenarios they can serve as templates for predictive hypotheses, once the relevant data has been gathered. In this capacity they can also serve as a guide to data collection, prioritising some data-gathering activities over others. In the case of the switch-trolley, we may change the way surveys are phrased to pick up on considerations of agency; in the case of funding simulations, we may look more closely at historical changes in research priorities following major discoveries.

Outside their predictive role, or lack thereof, the thought experiment and the simulation can help in other ways. Their main contribution is in clarifying connections between previously held beliefs. Given a domain about which we have various beliefs, drawn from various sources and held to different levels of confidence, the thought experiment or the simulation help us bring these beliefs into direct contact with each other, allowing the rejection of some and the generation of others. In other words, the concretisation allows a space for inference work. The arguments formulated in this inference work can then be transported to empirical contexts, where they can help interpret and infer from the data the relevant details and recommended actions.

5.6 Unrealistic assumptions

The simulation contains various unrealistic assumptions about the target domain. What effect does this have on the epistemic status of the results, and can these simplifications be overcome?

It is important to note here several simplifying, unrealistic assumptions involved in the simulation. While all simulations are likely to have some unrealistic assumptions, exploration of the unrealistic assumptions can suggest ways of extending the model. The very extendability, or openness, of the model is a virtue, as it allows model users to combine several models to fit a specific context of institutional design (see Chapter 2). In the current context, the simplifying assumptions do not detract from the main conclusion, that maximisation strategies under limited knowledge of dynamic processes would sometimes not work as well as random selection, because further complexities are only likely to further limit the ability of funding bodies to make good predictions about projects' future contribution to well-being.

5.6.1 Project length does not influence fitness contribution

In the current version of the simulation, when an agent's countdown reaches zero, the fitness contribution from that agent's work depends only on the agent's location, regardless of the time spent in that location. As the time spent corresponds to the duration of the project, we may expect that the magnitude of the contribution will be positively correlated with the duration of the project. This offers a possible extension of the model, in at least five ways:

1. Contribution multiplier: we can set the contribution of the agent to equal the fitness of the project multiplied by the duration of the project (perhaps with discounting), representing perhaps an elaborate packaging of the information in a way that makes it more fit.

2. Partial exploration: we can set the contribution of the agent to equal a fraction of the fitness of the project, with the size of the fraction increasing (perhaps with discounting) with the length of the project. This would represent partial exploration of the potentialities of the project.

3. Probabilistic success: we can set the contribution of the agent to be either zero or the fitness of the project, with a probability of success which depends on the length of the project. This would represent a certain threshold of difficulty which must be overcome in order to gain the results associated with the project. This method would be similar to the one used in the models presented by Kitcher (1993).

4. Local exploration: we can set the simulation such that upon completion of a project, the agent gains fitness from multiple locations in its local neighbourhood, correlated (perhaps with discounting) to the amount of time spent in the location. This would represent exploration of nearby projects during the period of the grant.

5. Combinations of the above.

5.6.2 Agents are identical

In the simulation, the agents are identical, in the sense that any agent, when placed at a location of a given fitness, will, when their countdown reaches zero, contribute that level of fitness. This

simplification ignores factors like individual natural ability and gained experience. These factors may be added to the simulation, especially in conjunction with the variations discussed above, as a differential propensity to generate fitness, which may be fixed (natural ability) or change over time (gained experience).

Overall, however, this point highlights the fact that the simulation presented in this chapter represents a particular approach to science funding, which focuses on funding *projects*, rather than funding *people* (see §2.1). The choice of this approach is informed by the explicit policies of science funding bodies, reflected, for example, in the institution of blind peer review. Nonetheless, a minority of science funding bodies, such as the Wellcome Trust, make explicit their preference to fund people rather than projects. I have discussed people-focused funding in Avin (2010), reaching similar pessimistic results about funding bodies' ability to gather enough information about the candidates to perform better than a lottery (an overview of the argument is given in the next chapter). However, that work has only been a preliminary study, and a full model, or suite of models, that combines both project merit and individual ability, would be informative for a more complete study of science funding.

5.6.3 Visibility is binary and single-valued

The simulation assigns binary values to the visibility of projects: either the community knows what the expected fitness contribution of a project will be, or it does not. A more realistic representation will allow partial visibility, with some distance decay effect, such that the community would still be able to make predictions of fitness contributions for less familiar projects, but these predictions will have a probability of being wrong, with the probability of error increasing the more unfamiliar (distant from explored locations) these projects are. This addition, however, will be computationally heavy, as it requires maintaining multiple versions of the landscape, both for the real values and for the estimated values. Taking this distance-decay to one possible extreme, it can be argued that peer review is less like the *best_visible* mechanism and more like *triage*, with projects that are dissimilar to past experience receiving effectively random evaluations but still having a chance of being funded. The random aspects already present in peer review are discussed in the next chapter.

A further simplification is in assigning only a single value of visibility. Even if we make the simplifying assumption that fitness is single-valued, there can still be many different *ex ante* estimations of the fitness of a proposed project, differing based on the individual visibility (knowledge of past and present projects) of the estimating agent. In real life, funding bodies address this problem by averaging the scores of several experts, each having a different experience of the scientific domain (see Chapter 1). Thus, we can think of the single-valued visibility map as the joint visibilities of the different experts. A more advanced model may simulate a network of experts, each connected to some subset of the agents exploring the domain. Such a model will allow mixing of the individual visibilities of the experts in non-trivial ways, e.g. allowing for dominance of some experts, or providing a score boost to candidates who operate in the networks of more than one expert.

5.6.4 Agents are non-interacting

The agents in the simulation are non-interacting, in the sense that their choice of location does not depend on the locations of other agents, both in the present and in the past, except indirectly via fitness dynamics and the visibility map. Agent interaction has been the central focus of most previous models of the social organisation of science, including Kitcher (1993); Strevens (2003); Weisberg and Muldoon (2009); Grim (2009). The omission of this interaction is not meant to reject its significance, which has been clearly demonstrated by these works, but rather to explore other important processes that have been neglected by these previous works, such as selection by funding bodies and fitness dynamics. To some extent, the processes of selection and dynamics included in the simulation mimic some of the processes which in the previous models are attributed to agent interaction. Nonetheless, the work leaves open the possibility of future extension which will include both interaction, selection, and fitness dynamics.

5.6.5 Smooth local fitness changes

In the current version of the simulation, fitness changes smoothly between adjacent projects. In part, this is a results of using bivariate Gaussians for hills of fitness, which influences the resulting ruggedness, as has been discussed in §5.3.1. However, another consideration regarding fitness is the interpretation of height as a linear scale, rather than, for example, a logarithmic scale. In keeping with Weisberg and Muldoon's model and fitness landscape models, the current version uses a linear scale for height. The use of a logarithmic scale instead of a linear scale would mean that the tops of hills are significantly (orders of magnitude) more fit than lower altitudes. The choice of scale depends on how one evaluates the achievements of science: if the history of science is read as a few revolutionary discoveries, interspersed with long periods of fruitless search, then a logarithmic scale may be more appropriate. Alternatively, if the history is considered as a gradual accumulation of minor successes, with the occasional jump that marks the opening of a new field, a linear scale would be more appropriate. The model could be modified to support a logarithmic scale, and results under such modification may prove interesting.

5.6.6 Inter-landscape interaction

The previous chapter listed various processes of fitness dynamics, only some of which were contained within a single corpus (and its associated landscape). However, for ease of calculation, the simulation in its current form only represents one landscape. Nonetheless, it is possible to some extent to represent many-landscape interactions without significant changes to the simulation, by simulating a larger configuration space, and mentally annotating different sub-spaces within it as different landscapes. The first two dynamic processes, *winner takes it all* and *reduced novelty* (§5.3.4), operate locally, and are therefore contained within each sub-space, but the *new avenues* dynamic, as well as the selection mechanisms, will operate across landscapes. We can therefore think of small landscapes as insular fields of research, whereas large landscapes represent multiple, theoretically or experimentally connected landscapes, with interdisciplinary fields in the intersections of sub-domains.

It is telling that already by including three types of dynamics we obtain significant results about the relative success of funding mechanisms in accumulating fitness. Of course, it would be interesting to explore simulations that include many landscape interactions, with a larger variety

of dynamic processes. In addition to multi-landscape simulation, it is possible, for example, to extend the simulation for other trigger types, e.g. following successful funding (seeding of new agents), or at a random interval during a project. It may also be interesting to assign effect values, either as bivariate Gaussians or other forms, which depend on the fitness of the triggering project or its duration. Finally, it might be interesting to explore an alternative where the location of the effect is not chosen purely at random across the entire landscape, but instead has a probabilistic location, with the probability featuring a distance-decay from the trigger location. All of these modifications could be used to create a more realistic representation of fitness dynamics, which will necessarily be more complex than the simple version presented in this chapter. For the current critique of peer-review, however, the simple dynamic picture presented is sufficient to support the sceptical argument.

Conclusion

The simulation succeeds in reproducing expected results for the easy cases for which we have strong intuitions, namely the behaviour on a static landscape with no selection, on a static landscape with selection, and the appearance of diminishing returns in long-term simulation. It also shows us that if we could know, *ex ante*, which are the most promising projects, regardless of whether we had experience of similar projects in the past, choosing projects based on that information would be the optimal science funding strategy. Unfortunately, this epistemic requirement is unlikely to be met in any real world situation.

For the harder case of short and medium term simulations on a dynamic landscape, the simulation clearly shows a significant effect for the inclusion of fitness dynamics on the process of fitness accumulation. Interestingly, it shows that on small landscapes random selection of agents fares roughly equally as selecting the most promising projects from those with which we have some familiarity, and roughly equally as performing no selection, keeping the same agents going in long research programmes. More interestingly, on larger landscapes random selection significantly outperforms the two alternatives mentioned. As discussed above, larger landscapes can be taken to represent highly inter-related areas of research, whether by theory or by practice. The implications of these results for the design of an alternative science funding mechanism are discussed in the next chapter.

Chapter 6

Funding Science by Lottery

Let us be honest
about the random aspect
of science funding.

Introduction

In the previous chapter an interesting result has been reached: in some cases, using a lottery to choose which projects to fund performs better than picking the best out of those projects with which the scientific community has some familiarity.[1] This result was achieved by investigating a social model of scientific progress, which has been developed as a revised version of earlier models by philosophers of science (these earlier models were discussed in Chapter 2). The measure of performance used is the accumulation of epistemic fitness, which is the aim of public science funding bodies (Chapter 3). A key novel assumption in the model is that epistemic fitness changes over time, as a response to the actions of scientific investigators (Chapter 4). Modestly generalised, the conclusion of the modelling work in the previous chapter is that current funding agencies are *not* pursuing the most rational funding strategy; funding by peer review seems, according to the model, analogous to the proverbial man searching for his keys under the lamp-post because "that is where the light is".

The model has also been used to show that if we could somehow know the fitness of unexplored domains, choosing by that fitness would prove most beneficial. This led to the simulation of a half-way strategy labelled "triage", which involved a combination of optimising based on what is known, and allowing room for (random) exploration of the unknown. This chapter is dedicated to the exploration of how an efficient balance may be achieved between optimisation and exploration, by using some combination of merit evaluation and a lottery, in a real-world science funding mechanism.

The target in designing this mechanism is not utopian, but realistic: I'm presenting here a template for an actual science funding institution. It need not be optimal, merely better than the current system we have in place, namely funding by ranking based on peer review (Chapter 1). The starting position of this chapter is that the models have shown that (some) funding by lottery can be more *effective* than funding by peer review (according to the set of desiderata for a funding mechanism presented in §1.4). However, when it comes to the design of a real-world

[1] A proposal to fund science by lottery was recently made by Gillies (2014), as discussed in Chapter 1.

mechanism, we need to go beyond the models: we need to include more empirical data, and we need to consider desiderata other than effectiveness. The early sections of this chapter (§6.1 and §6.2) complement the argument of the previous chapters with empirical evidence and broader theoretical considerations. This further evidence is used in the design of a specific funding mechanism, presented in §6.3. Finally, some limitations of the proposal are discussed in §6.4.

6.1 Empirical evidence for problems with allocation by peer review

The first step in bringing the lottery proposal into the context of contemporary science policy is to ask what problems with the current system (which, at least as presented, is not a lottery) the lottery may solve. As discussed in Chapter 1, the current dominant system of public science funding is peer review, where proposals are invited from practising scientists, and these proposals are then evaluated by peer review for merit. Funding is allocated according to this peer evaluation, from the most meritorious downwards until the funds run out. Opinions about the merits of the peer review system, and its shortcomings, are numerous and varied.[2] Empirical evaluations of aspects of the system are more rare (Demicheli and Di Pietrantonj, 2007), but stand to provide a clearer insight into what might be deficient in the peer review system, and where introduction of random elements in the vein of the lottery proposal may improve the system by simultaneously increasing the overall fitness of projects selected and by reducing the cost of operating the funding mechanism. Two such studies are presented below: the first looks at the level of randomness already present in the peer review system; the second looks at the cost of running the peer review evaluation.

6.1.1 Measuring the variability of peer review scores

How can we measure the effectiveness of peer review? One fairly good measure would be to compare the scores of reviewers to the actual impact of funded projects. Such a measurement would give us an estimate of the validity of the merit scores assigned by reviewers. However, the ability to conduct such studies is very limited. For example, Dinges (2005) conducted an evaluation study of the Austrian science fund (FWF), using data gathered by FWF regarding funded projects, including publication record, employment of researchers and staff, and an *ex post* evaluation of the projects by anonymous peers. While Dinges provides useful insights for the researchers and FWF staff involved in the funding exercise, he is also very explicit about the limitations of this kind of study:

- Information is only available about funded projects. Thus, there is no way of evaluating whether the system is effective at funding the *best* proposals, only the extent to which funding the chosen projects produced a benefit. Thus, it cannot help chose between substantially different methods of funding; at best, it can provide justification for having public funding of science *at all*, and perhaps propose small tweaks to the current system.
- The *ex post* evaluations of projects' success and impacts were carried out by the same experts who evaluated the project proposals and who contributed to the funding decisions,

[2] For a positive evaluation see Polanyi (1962); Frazier (1987); Research Councils UK (2006). For criticisms see Chubin and Hackett (1990); Martino (1992); Gillies (2008, 2014). See detailed discussion in Chapter 1.

which is likely to lead to significant positive bias.

- Measurements of publications and citations (bibliometrics) are poor indicators when applied across multiple disciplines and fields, as publication and citation practices vary significantly. As discussed in Chapter 3, public science funding bodies often support a range of disciplines, or large heterogeneous disciplines, and so direct use of metrics in *ex post* evaluation would prove tricky.

- There are no established indicators for measuring the impact of science. The indicators that exist in the literature are dominantly economic, and are ill-suited to measuring the impact of basic research. In a table adapted from Godin and Doré (2004), Dinges (pp. 20-21) lists 61 different types of possible indicators for scientific impact, the majority of which are not currently measured. Furthermore, problems of operationalisation and measurement are likely to be present for many of the proposed indicators, due to their intangible or subjective nature.

The above list is not exhaustive, but it is sufficient for establishing the difficulty, at least at present, of directly measuring the effectiveness of funding methods in generating positive societal impacts, and the related difficulty of comparing alternative funding methods with regards to their primary function of choosing the best research.

A weaker evaluation of the validity of the scores of peer review is to check their consistency: to what extent different panel members agree among themselves about the merit of individual projects. Such a measurement is clearly more limited in what it tells us about the reliability of peer review. Assume (unrealistically) that there is some true measure of the merit or fitness of a proposed project in the same way there is a true measure of the length of a stick, neglecting for now the inherent value-laden and dynamic aspects of fitness. We can then treat each reviewer's evaluation as an estimate of that measure, with some possible random error and some possible bias, as if each reviewer's assessment is analogous to an independent measurement with a different ruler. Since there is no external measure of project merit or fitness, as discussed above, we can never rule out the possibility that a systematic bias is operating on all reviewers, such that close agreement between reviewers is no guarantee of a reliable measure (all our rulers might be wrongly marked in the same way). A wide spread of scores, while telling us nothing about bias, will give us an indication that each individual estimate is subject to large variability (we will know that something is amiss with our rulers if consecutive measurements yield very different results). In the case of peer assessment, we can hypothesise that the source of any observed variability is due either to objective uncertainty, objective differences between reviewers' experience, or subjective differences between reviewers' interests and values. In this scenario of a simple measurement, increasing the number of estimates will increase the reliability of the mean. Therefore, an estimate of variability will indicate the number of reviewers required to make a reliable estimate of the merit of each project. Alternatively, the variability can indicate the level of (un)reliability (only due to error, not bias) of mean scores given a certain number of reviewers.

The most thorough measurement published to date of the variability of grant peer review scores was conducted by Graves et al. (2011).[3] The authors used the raw peer review scores

[3] An earlier review paper by Cicchetti (1991) covers various measurements with smaller sample sizes. The paper, published alongside insightful reviewers' comments, is rich in discussion of the evidence available at the time, and the statistical tools suitable for this kind of measurement.

assigned by individual panel members to 2705 grant proposals. All proposals were submitted to the National Health and Medical Research Council of Australia (NHMRC) in 2009. The scores were given by reviewers sitting on panels of seven, nine, or eleven members, and the average score of the panel was used to decide whether a project was funded or not, based on its rank relative to other proposals.

The authors used a bootstrap method to obtain an estimate of variability of the mean of peer review scores from the available raw scores.[4] In this method, a set of bootstrap samples, often 1,000-10,000, are obtained from the original sample (in this case, the raw scores of a single proposal), by randomly selecting scores from the original raw scores *with repetition*, until a set of the same size is obtained. For example, if an original set of raw scores was $\{3, 3, 4, 4, 6, 7, 9\}$, giving an average of 5.14, one of the bootstrap samples might be $\{3, 4, 4, 4, 6, 9, 9\}$, giving an average of 5.57, but not $\{3, 4, 5, 6, 7, 8, 9\}$, as 5 and 8 did not appear in the original panel scores. Due to the random sampling, the likelihood of any score appearing in a bootstrap sample is related to the number of appearances it had in the original panel, so in the example above any individual score in any bootstrap sample is twice as likely to be 3 or 4 than 6, 7 or 9. The set of bootstrap samples is then used as a proxy for the population of possible samples, yielding a mean and a variance in that mean, and a confidence interval around the mean. This confidence interval, labeled by the authors the "score interval", was then compared to the funding cutoff line: proposals whose score interval was consistently above or consistently below the funding line were considered "efficiently classified" by the review system, whereas proposals whose score interval straddled the funding line were considered as problematic, or "variably/randomly classified". A bootstrap method was chosen because the sample sizes are small, prohibiting the use of more direct estimations of variability, and because the underlying distribution of potential review scores is unknown, and cannot be assumed to be Gaussian.

The results of this bootstrap method showed that overall, 61% of proposals were never funded (score interval was consistently below the funding line), 9% were always funded (score interval consistently above the funding line), and 29% were sometimes funded (score interval straddling the funding line). The results also showed variability between panels (which correspond to different scientific areas), an effect which may be explained using the models developed in the previous chapters, e.g. by showing that different panels dealt with epistemic landscapes of different sizes, as the results from the last chapter show that the relative effectiveness of peer review depends on the landscape size. More information, however, is required to check whether this explanation agrees with empirical data, and this is not pursued further in this thesis.

In the authors' opinion, the discrepancy between the observed levels of variability, and the importance of funding decisions to individuals' careers, is cause for concern. The authors claim the results show "a high degree of randomness", with "relatively poor reliability in scoring" (p. 3). As the authors are themselves practising scientists, who participate in the grant peer review system as both applicants and reviewers, we cannot discard their views off-hand. The authors follow with a list of possible improvements to the peer review system. One of their suggestions is to investigate the use of a (limited) lottery:

> Another avenue for investigation would be to assess the formal inclusion of randomness. There may be merit in allowing panels to classify grants into three categories: certain funding, certain rejection, or funding based on a random draw for proposals that are

[4] For an introduction to bootstrap methods see Davison and Hinkley (1997).

difficult to discriminate. (Graves et al., 2011, p. 4)

Despite their concern, the authors do not offer a hypothesis for the origin of high variability (though a later paper, discussed below, does offer such a hypothesis). Given the modelling work of earlier chapters, one reasonable explanation would be that the variable scores are assigned to proposals outside of the "vision range" of reviewers. There is no way to check this hypothesis without access to the contents of the proposals and the expertise of the reviewers. Other possible explanations would be that reviewers have varying subjective preferences with which they evaluate proposals, or different views of the relevant scientific discipline which they were not able to commensurate while on the panel, or that reviewers vary in their ability (cognitive or other) to evaluate the merit of a project given a written description and a knowledge of the scientific discipline.

The above quote suggests the authors see a link between variability in scores and a (limited) use of a lottery in funding. While this is not the line taken by the authors, this link can be made even more suggestive, if we think of the workings of current funding panels *as if* they were an implementation of the system described in the quote. If we black box the workings of the panel, and just look at the inputs and outputs, we see 100% of the applications coming in, the top 10% or so coming out as "effectively" funded, the lower half or so being "effectively" rejected, and the middle group being subjected to some semi-random process. Even if we look into the black box, we can see that the process of expert deliberation, when applied to the middle group, bears strong resemblance to the process of a random number generator: it is highly variable and largely unpredictable. Specifically, and importantly, the psychological and social deliberation process for the middle group resembles the operation of a "true" or "physical" random number generator, such as a lottery ball machine or a quantum measurement. In such a setup, the unpredictability of the mechanism is due to high complexity or an inherent unknowable nature of the system.[5]

Thus, we could conclude that funding by peer review *is* funding by triage, with random allocation for the middle group. However, there are three distinct differences between peer review and triage with formal randomness: the cost of the operation, the appearance of randomness, and the agency of the reviewers. These three differences are discussed below and later in the chapter.

6.1.2 Measuring the cost of grant peer review

The cost of the grant peer review system can be broken down into three components:

1. The cost of writing the applications (both successful and unsuccessful), incurred by the applicants.

2. The cost of evaluating the proposals and deciding on which application to fund, incurred by internal and external reviewers.

3. The administrative costs of the process, incurred by the funding body.

According to Graves et al. (2011), in the funding exercise discussed above the largest of these costs was, by far, the cost incurred by the applicants, totalling 85% of the total cost of the

[5]These random generators are different from pseudorandom number generators, such as the algorithms in operation in computers and pocket calculators, which rely on well-studied mathematical systems that guarantee high variability and equal chances to all possible outcomes. For an introduction to random number generators see Knuth (1997, Vol. 2).

exercise (p. 3). The authors used full costing of the review process and administration budget, but only a small sample of applicant reports. To complete their data, a more comprehensive survey was conducted amongst the researchers who submitted applications to NHMRC in March, 2012. The results of this survey, discussed below, are reported in Herbert et al. (2013).

The authors received responses from 285 scientists who submitted in total 632 proposals. These provide a representative sample of the 3570 proposals sent to NHMRC in March 2012, and display the same success rate of 21%. Based on the survey results the authors estimated, with a high degree of confidence, that 550 working years went into writing the proposals for the March 2012 funding round. When monetised based on the researchers' salaries, this is equivalent to 14% of the funding budget of NHMRC. New proposals took on average 38 days to prepare, and resubmissions took on average 28 days. The average length of a proposal was 80-120 pages.

Surprisingly, extra time spent on a proposal did not increase its probability of success. Neither did the researcher's salary, which is an indicator of seniority. The researcher's previous experience with the peer review system, either as an applicant or as a reviewer, did increase the probability of success, though the increase was not statistically significant. The researchers' own evaluation of which of their proposals would be funded bore no significant correlation to the actual funding decisions. The only statistically significant effect on probability of success was that resubmitted (failed) proposals were significantly less likely to be funded, when compared to new proposals.[6]

The authors' recommendations are largely unsurprising given the findings: time wasted should be reduced by having multiple funding rounds with increasing information requirements, and there should be an exclusion period for failed applications before they can be resubmitted. What is more interesting is the authors' conceptualisation of their findings. The authors hypothesise the existence of a curve which associates the accuracy of the peer review system in evaluating the merit of a proposal to the amount of information provided by each applicant (Fig. 6.1, in black). Such considerations of accuracy in measurement of merit are similar to the notion of fitness evaluation discussed in the previous chapter.

The hypothetical graph of Herbert et al. has certain interesting features:

- The graph hypothesises the existence of an "ideal", which is the amount of information required for the optimal level of accuracy. In the paper this level of accuracy appears close to, though not equal to, 100%.

- In the area left of the "ideal", i.e. where the information provided is less than the ideal amount, the graph displays diminishing returns, such that equal increases in information provided result in less increase in accuracy the more information has already been provided.

- In the area right of the "ideal", the graph displays an "overshoot" effect, with accuracy decreasing as information increases. In the text, this is explained as the reviewers being overburdened with too much information.

The authors rely on their finding, that extra time spent on a proposal does not increase its likelihood of success, to argue that the current amount of information provided is more than

[6]The authors do not provide a hypothesis to account for this statistically significant observation. Given the modelling work in the previous chapters, we could hypothesise that a significant portion of failed proposals represent low fitness projects within the visibility range of the scientific discipline. Since, over a short period of time, significant gain of fitness is more rare than significant loss of fitness (as the field progresses it "exhausts" the area of familiar projects), what is once labelled as low quality (if within the vision range) is likely to be similarly labelled in subsequent years, until a rare long-range effect re-infuses the exhausted area with new fitness.

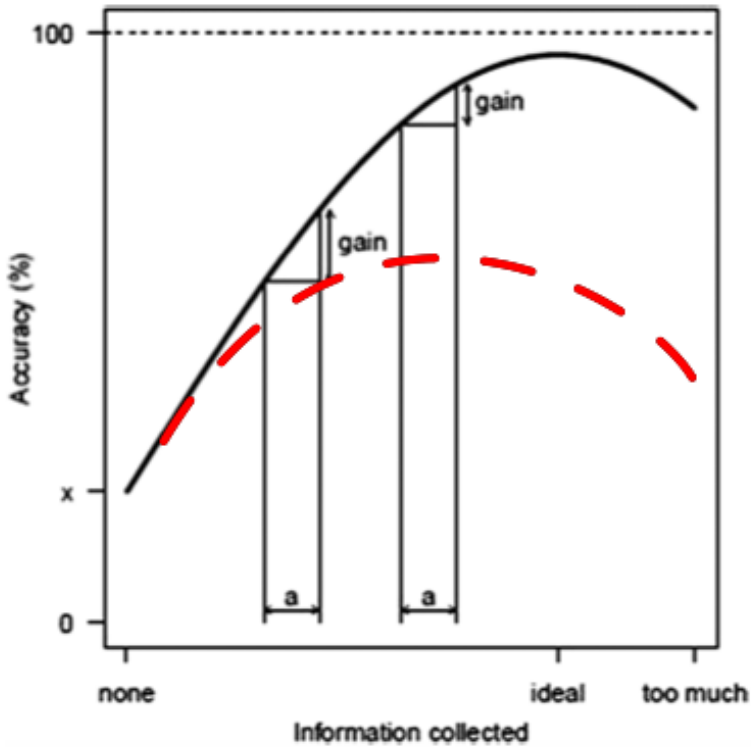


Figure 6.1: The accuracy of peer review assessment as a function of information provided. Original figure, in black, is reproduced from Herbert et al. (2013, Fig. 2, p. 5), and represents the authors' hypothesis, not a conclusion from their data. The red dashed curve was added by me, and represents an alternative dependance, based on the finding of Chapter 5. Herbert et al. (2013) was published under CC-BY-NC licence: `http://creativecommons.org/licenses/by-nc/3.0/legalcode`.

the ideal. However, one does not follow the other, because increased accuracy does not imply higher merit for a proposal. Nonetheless, the authors' description of reviewers having to read 50-100 proposals of 80-120 pages does suggest an unnecessary cognitive burden. Based on their hypothetical curve, the authors' suggestions for reducing the amount of information gathered implies a lower accuracy for the peer review system. The authors believe this lowered accuracy is justified, on cost/benefit grounds, even though in their model a high level of accuracy is possible. However, given the sceptical arguments in the previous chapters, it is quite possible that a high level of accuracy is not even possible, and therefore requiring scientists to provide less information is not only an efficient compromise, it is in fact epistemically optimal (Fig. 6.1, dashed red curve).

The empirical studies discussed above show that despite high costs, the peer review system leaves an epistemic gap between the information provided in the proposals and the genuine fitness of projects, such that high variability exists for a significant middle group. This result was expected given the model of the previous chapter, and suggests, using the language of the model, that scientists are not completely discouraged from submitting proposals that lie outside the vision range gained from past projects. In the model, access to areas outside the vision range was gained by a lottery. While the empirical results suggest that access to unexplored areas can be achieved by peer review as well, the use of a lottery for treating the high-variability middle group may have other advantages, for example in cost reduction. These advantages are discussed in the following section, which presents two theoretical studies on the use of lotteries.

6.2 Theoretical background on lotteries

Lotteries have been used in the past, and in some cases are still being used, for distributing various goods, such as the right to rule, money prizes, hunting permits, admittance to sought-after schools and university courses, citizenship, and many more, as well as various "bads", such as military draft or jury duty.[7] The prevalence of lotteries and their unique features have generated various theoretical works in political theory, economics, and moral philosophy.[8]

This section presents two theoretical investigations of the use of lotteries for cases which bear some, though only partial, similarity to the case of science funding. Partial similarities would have to suffice, as there has been no comprehensive theoretical study on the use of lotteries for science funding. The studies help highlight elements of distribution by lottery, and distribution in general, which have been missing from the models presented in the previous chapters.

Boyle (1998) discusses the use of lotteries for the selection of individuals to a coveted post, following an initial aptitude test. Boyle highlights the importance of the appearance of fairness in a selection process that selects individuals. He makes the case for using a lottery even when a fairly good, though not perfect, predictor of success is available, and so his argument would apply even more forcefully to the science funding case, where we have reasons to doubt the existence of such a predictor (or at least our ability to access it).

Boyce (1994) considers the economics of distribution by lottery, as opposed to other distribution mechanisms such as merit evaluation or auction. While even further from the case

[7] A comprehensive and well-researched list of current and past lotteries is available on the website of the Kleroterians, a society of scholars advocating the exploration of the use of lotteries (Boyle, 2013).

[8] Books of note on the topic of lotteries include Boyle (2010); Duxbury (1999); Gataker and Boyle (2008); Goodwin (2005); Stone (2011).

at hand, as it deals primarily with goods such as hunting permits designed for consumption, Boyce's account highlights two important issues which have been left out of the model of the previous chapter: the cost of running the distribution mechanism, and the expected behaviour of agents participating in such a system. The paper can be used, by analogy, to argue that distribution by lottery would be cheaper than a costly merit evaluation, and may also help deflate unrealistic promises made in project proposals. Such deflations can help make science funding more accountable to the public, and also improve its perception amongst participating scientists (see the list of desiderata in §1.4).

The two papers also highlight an important aspect of science funding missing from the models, that of difference in ability in the cohort of applicants. This aspect will be addressed later, in the section presenting the design of the proposed funding mechanism.

6.2.1 Introducing lotteries to selection mechanisms of individuals by organisations

Boyle (1998) proposed, in a paper presented to the Royal Society of Statisticians, that graduated lotteries be introduced into processes where individuals are selected by organisations based on fallible measurement criteria, in order to increase the fairness of the process without significant loss of efficiency. Boyle develops this idea from the Victorian economist and statistician Edgeworth (1888, 1890), who in a couple of papers discussed the random element in the allocation of grades in university exams, and the potential benefit of introducing a weighted lottery based on the results of a "light" examination (of an unspecified nature) in the selection of candidates to civil service positions, instead of using the results of university exams. It is assumed that the exam cannot be improved, or if the exam is improved, its best form will still involve some residual random element. In Edgeworth's proposal, students just above and below the cutoff line will be given a number of lottery tickets corresponding to the probability that they deservedly belong above the cutoff line, based on the estimated error in the light exam. According to Edgeworth, the replacement of the fine-grained examination with such a weighted lottery would not significantly decrease (in the long run) the amount of good candidates being admitted to the program, and further it would have two benefits:

1. It would mitigate the sense of injustice felt by those candidates who, under the examination method, would score just under the cutoff line.

2. It would alert the public to the random component of examination scores.

Boyle develops and refines Edgeworth's proposal in a series of steps. The first step is to consider in some detail two desiderata of selection mechanisms: efficiency and fairness. These are also key desiderata for a science funding mechanism, as discussed in §1.4. Boyle's definitions do not map exactly to the definitions of efficiency and fairness given in §1.4, and are therefore presented here:[9]

[9] As was mentioned in §1.4, there is not always a clear distinction between considerations that relate to fairness and considerations that relate to efficiency. For example, avoiding factors that are irrelevant to a candidate's performance, such as their gender, could be labelled both as a fairness consideration and as an efficiency consideration, as it would be inefficient to rule out good candidates based on irrelevant factors. Since this section argues that lotteries are preferable to more direct evaluations *both* on efficiency grounds *and* on fairness grounds, making this distinction clear at every instance is not very important.

Efficiency At its simplest form, efficiency is the achievement of maximal beneficial outcome for minimal cost. Boyle gives an example of reducing post-natal infant mortality (Carpenter, 1983): the health organisations measured various indicators of infant risk, combined them to a single measure, and directed extra care to those infants who scored above the "care line". This policy successfully reduced infant mortality rates, and can therefore count as *efficient*.

Fairness Boyle, while admitting the complexity of the concept of fairness, adopts Elster's working definition of fairness, of treating relevantly like cases alike (Elster, 1989). Boyle, following Elster, elaborates four criteria for fairness in the selection of people:

1. The selection process should minimise wasted effort by applicants, e.g. by not requiring information which is superfluous or irrelevant, by not demanding extensive travel etc.
2. The selection process should not make a clear cutoff between candidates whose measurable difference is not statistically significant, e.g. due to random error in measurement scores.
3. The selection process should avoid bias, both intentional and unintentional, e.g. sexism or racism, but also "heightism" or "hairism".
4. The selection process should be free from corruption.

Note that none of these criteria relate to relevant differences; According to Boyle's account, a system which treats all candidates exactly alike would be considered *fair*, though it will probably be *inefficient*. For example, under Boyle's account, if candidate A has some demonstrable and relevant qualities that are better than candidate B's, but A failed to score significantly higher than B on the chosen test (which assumedly checks for these, and other, qualities), it would not be *unfair* if B is consequently picked for the position instead of A, though it might have been more *efficient* if A was picked instead of B.

While the drive for efficiency is often internal to the organisation, there are often external drivers for fairness, including laws (e.g. against discrimination), and public scrutiny of selection results (either via high profile cases or via published statistics). In the case of science funding it seems the drive for efficiency would also be external, e.g. from Congress in the case of US funding bodies. It seems reasonable to generalise here and say that when individuals are selected for some productive roles, the issue of fairness will be of concern among the population applying for these roles (and their extended social circle) and the issue of efficiency will be of concern to those how are positioned to benefit from the products of labour. In Boyle's case the products of labour are enjoyed by the organisation performing the selection, whereas in the (public) science funding case the products of labour are enjoyed by society (as discussed in Chapter 3).

Boyle proposes the following example of how a lottery might have been introduced into a selection mechanism to make it more fair.[10] In the old British grammar school system, an IQ test, called the eleven plus test, was given to students at age eleven, and the high scorers in each local education authority would be given places in the more academically-oriented grammar schools.[11] The eleven plus IQ test was considered the most reliable predictor of the five-year

[10]The example's link to science funding is tenuous, but it provides material for Boyle's generalisations that follow it. These generalisations have more direct relevance to the thesis.

[11]In some places in the UK the exam is still in use, though mostly as a voluntary exam taken by pupils who seek admission to a particular group of schools. Boyle's paper addresses the exam in the context of the old system, where it was used as a blanket exam for all students seeking admission to a grammar school in the UK.

academic success of students out of the available measures, though it was known not to be perfect. Initially, a "border zone" near the cutoff score for admittance was created, and children who scored in the "border zone" were further evaluated using teacher reports and other information. Over time, probably for administrative reasons, the border zone was shrunk. Boyle claims that the border zone should not have been shrunk, and if anything, it should have been expanded. He claims the border zone should be set according to the possible error in the test: marking errors account for 1% error rate, repeatability errors (children's performance varying on different sittings) account for 10% error rate, and prediction errors (the test not correctly predicting academic performance) account for 15% error rate, and in total Boyle arrives at a 26% error rate. Given a normal distribution of results, and admittance rates to grammar schools of 25%, this yields a "border zone" of 40% of students, those who scored in the top 45% but excluding the top 5%.[12] From this, Boyle suggests the following:

1. Automatically admit the top 5%, who performed significantly better than the other candidates.

2. Automatically reject the bottom 55% percent, who performed significantly worse than the other candidates, and where there is a very small chance they scored below the cutoff line by mistake.

3. For the remaining 40%, perform a "graduated" lottery, as depicted in Fig. 6.2, such that 3/4 of the lowest 10% are chosen at random and joined with the second-lowest 10%, from these 3/4 are chosen and joined with the second-highest 10%, and so forth until in the end only half the candidates remain, forming 20% of the original population, and together with the 5% who were selected automatically they form the admittance quota of 25%.

According to Boyle, this mechanism will have the following advantages:

1. A lottery is quick, cheap, and random, reducing both the direct cost to the applicant (compared with, say, more testing) and the indirect costs by reducing the incentive to spend extra effort on the test (i.e. reduce the motivation to slightly exaggerate one's own abilities).

2. From the point of view of the candidates, a lottery is fairer, as it treats those who are not distinguishable in a statistically significant manner as the same.

3. While no process could be completely free from bias, a lottery gives every candidate, whatever their public standing, a non-zero, measurable chance of success. This is true regardless of any particular anti-bias mechanisms that are in fashion at the time.

4. A publicly visible lottery is, to a large extent, free from corruption, as no individual has power over the direct outcome. Bureaucrats without taint of corruption may be even better, but they are hard to come by and expensive to maintain.

5. A lottery could reduce the costs the organisation spends on proving to external parties the selection mechanism is fair.

[12]The similarity between Boyle's numbers and the numbers of Graves et al. is largely accidental, arising mostly from the similar arbitrary cutoff percentages of 25% and 21%, respectively. Nonetheless, the similarity is convenient for translating, at least as a mental exercise, from one context to the other.

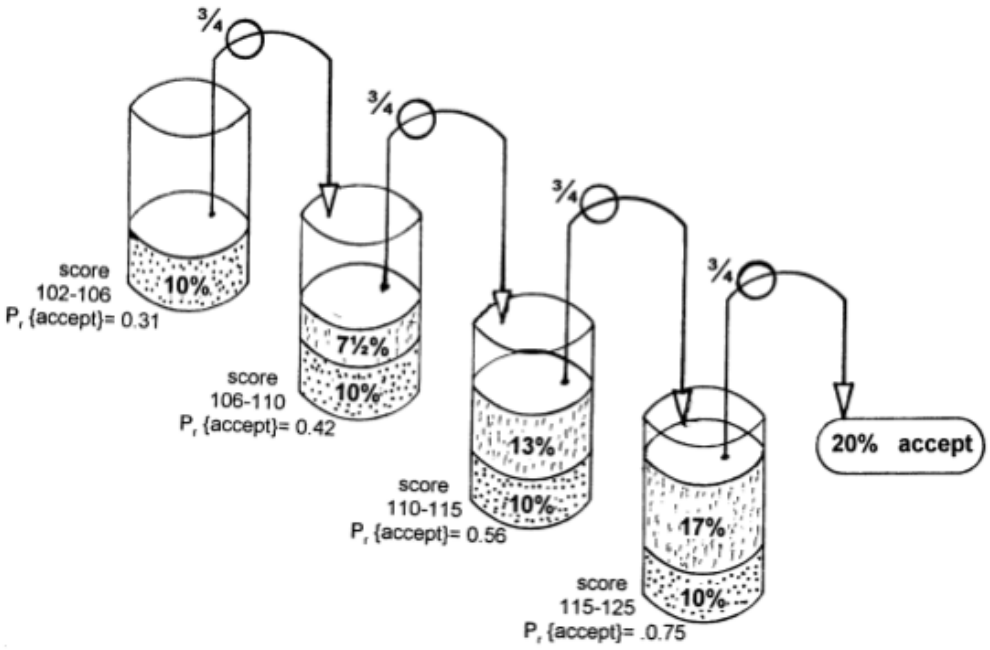


Figure 6.2: A schematic of a graduated lottery, from Boyle (1998, Fig. 2, p. 301). Permission to reproduce was obtained from the publisher under licence number 3399410894580.

6. A lottery may benefit the organisation by occasionally introducing into the selection pool candidates who have rare and valuable skills which are not picked up by the test.

Boyle's lottery and science funding

Boyle's argument can be applied, with some modification, to the context of project selection for science funding, though some key differences must be remembered:

1. In the science funding scenario the selection is among project proposals, not people. Nonetheless, the decision does directly influence the lives of the researchers associated with each proposal, and so considerations of fairness and psychological effect on participants have their place.

2. If we adopt a society-wide perspective (as we did in Chapter 3; more on this later), it is both more efficient and more fair to pick the projects of highest fitness, because fitness takes into account the information needs of the entire population. Nonetheless, when comparing mechanisms of equal ability to gain fitness, the mechanism that is more fair on the participating scientists would be preferred.

3. There is currently no good estimate of the predictive power, and the related error or uncertainty, of the proposal evaluation process, though the arguments in the previous chapters suggest it will be large. Following the arguments in the previous chapters, a significant portion of the error or uncertainty in evaluating proposals may be ineliminable, because the information required simply does not exist at the time of evaluation. Nonetheless, we can use the measurements of variability discussed in §6.1 as a guideline for setting up the "border zone" for grant proposals.

Using the model presented in the previous chapter, we can draw up a three-group distinction similar to that drawn by Boyle:

1. Projects of high fitness within the "vision range" of expert assessors, i.e. projects which are similar to ones which have previously yielded high fitness results, and where it is known that the potential for advancement is far from being exhausted, e.g. elucidating and expanding studies directly after a significant breakthrough. These should be funded directly.

2. Projects of medium fitness within the "vision range", and projects outside the "vision range". The two groups are brought together, even though they are expected to play out very differently, in order to allow a wider set of approaches representation in the research portfolio, increasing the chances of rare breakthroughs, and also the chances of meeting the needs of minority or under-represented groups, while accepting the cost of admitting from among the projects outside the vision range ones of low fitness. As the number of these projects is likely to exceed available funding, selection should be made by a lottery, possibly graduated based on peer review scores.

3. Projects of low fitness within the "vision range" should be rejected. These would include, for example, projects which directly replicate (e.g out of ignorance) past projects, projects which follow a long-deserted research programme without good justification for its revival, or projects whose design fails to appreciate or follow well-known standards and best practices.

A similar proposal, though less elaborate and with less supporting justification, was presented by Greenberg (1998). A more detailed version of the above, and a discussion thereof along the theoretical lines sketched by Boyle, is presented later in the chapter. However, we have yet to consider the validity and strength of Boyle's argument and his proposal. Boyle's paper was published alongside comments from various experts, including moral philosophers, statisticians, an occupational psychologist responsible for entry examination tests, an administrator of school examinations, a marketing expert, and an insurance expert. Some of the comments apply to the science funding case, while others do not. A summary of the relevant comments is provided below, alongside Boyle's responses to the criticisms.

Criticisms and responses to Boyle's paper

A common criticism, both from statisticians and examination administrators, was that a lottery would more often substitute a truly meritorious applicant with a less meritorious applicant than would a test. This was considered an important shortcoming in efficiency, but also considered to be unjust from the point of view of the more meritorious applicant. The statistical details of this argument were in effect identical between the commentators, and can be exemplified in the following model: label the real value, which precisely predicts the performance of candidate i, as T_i, and the test result score for that candidate as t_i. The error in the test for that candidate is then $e_i = T_i - t_i$. For a well-designed test, this error will be *random* rather then *systematic*, which means it will be normally distributed around a mean value of 0.[13] Thus, if we compare two candidates, the error in the test would equally apply to both, and the likelihood that the higher scoring candidate will be the better achieving one is greater. The outcome of the test may not be fair, as the test results of one candidate may be higher than the results of another candidate

[13]When the long term achievements of candidates are measurable, as in the case of the IQ test and academic achievement, the tests can be tested for systematic errors, and correction mechanisms which may include some randomisation are sometimes included, e.g. in the order of the questions. As discussed earlier in the chapter, no good mechanisms for empirically uncovering systemic bias in peer review results are available.

of equal-merit, and lead to the first candidate getting the job; however, both candidates were admitted to the same process, and were equally subjected to the same probability of error. The potential error in the test in fact serves as a kind of lottery, which operates on top of the main function of the test, which is to predict performance.

Boyle responds to this criticism by first agreeing that merely adding a purely random score to the test scores of candidates would serve no beneficial purpose. However, he defends the graduated lottery on three grounds:

Non-linearity The criticism assumes that higher test scores correspond to higher achievements throughout the range of scores, i.e. that the test score is linearly dependant on the real value. However, Boyle claims, there is evidence that, for example in the case of IQ, beyond a certain threshold higher scores no longer predict higher achievement, even if the test succeeds in making predictions for lower scores. Thus, even if the test is reliable when the entire range is considered, if the cutoff score is higher than or near to the point of non-linearity, the criticism no longer holds, since within the new border area the test is no longer a good differentiator of candidates.

In the science funding case, unlike the case of IQ tests, there is no evidence of reliability for any range of scores, and so worries regarding non-linearity are expected to be even more relevant.

Systematic bias Boyle argues that the test is likely to be designed to pick up a few traits which are strongly correlated with success, while ignoring a range of other, more rare or difficult to measure traits. This introduces two possible sources of systematic bias, which, if not directly controlled for, could undermine the efficiency argument:

- The key traits tested for may be more easily detected in a certain subset of the population, leading to unfair treatment by the test, e.g. logic questions relying on a certain level of linguistic comprehension which favours native speakers even if the job does not require language skills. As mentioned, effective comparison of test results with later performance can help screen for such bias, but only if such comparison is carried out in an effective manner, and if the measures of performance themselves are free of bias.

  As discussed above, there are at present no good measures for eliminating systematic bias from grant peer review, because there are no good *ex post* indicators, and because no data could be had on the success of unfunded projects (as opposed of the academic success of children who went to less-academic schools). As mentioned in §1.4, studies measuring the performance of particular minority groups in grant peer review do exist, and detected biases sometimes lead to the establishment of dedicated funding pools, though this tends to be very controversial (Chubin and Hackett, 1990; Chubin, 1994).
- The unmeasured traits which can lead to success may be negatively correlated to the measured traits, e.g. if a deficiency in a key trait provides the necessary motivation to develop rare skills. For example, creative "out of the box" thinking, which can be valuable in certain problem-solving situations, is often suppressed among individuals who are very proficient in specific analytic, semi-algorithmic problem solving skills. A test for the latter kind of skills will be biased against those candidates who are strong in the former set.

Similarly, in the science case, highly innovative thinking may be correlated to low evaluation based on the prevailing "paradigm", as argued by Gillies (see §1.6).

In both cases of bias, the criticism that tests are better than lotteries at selecting the best candidates is undermined because we have reason to suspect that the "error" in the test is not normally distributed for all individuals in the population, the test is therefore not an "effective lottery", and its claim to fair treatment of all candidates is undermined. More blatant cases of bias could also be counted here, such as bribery and overt racism and sexism (as opposed to hidden biases that result from the choice of evaluative criteria).

Diversity As mentioned in some of the comments on the paper, one of the possible advantages of a lottery over a test is to promote diversity, by preventing "cloning" of existing candidates. This is not a comment about fairness, but a comment about efficiency: it is better *for the organisation* to have a more diverse workforce, to allow diverse thinking and learning. This efficiency consideration, which takes into account the cohort of recruits as a whole, is different from the efficiency consideration of the test, which is only a measure of how well the test predicts the performance of individual candidates and supports good selection decisions based on these individual predictions. Thus, the argument goes, to maximise efficiency it is good to have mechanisms that address both aspects of efficiency (individual-level and group-level), and a lottery serves group-level efficiency better than a test would, by increasing diversity.

This argument by Boyle is directly supported by the modelling work presented in the previous chapter, and bears very strong resemblance to Gillies' argument against the homogeneity-inducing effects of peer review (see §1.6).

Another criticism, presented by Goodwin, argued that by the logic of the argument, and given the long tail of error distributions, *all* applicants should be admitted to a graduated lottery. This argument is a local and restricted version of Goodwin's more general advocacy for the use of lotteries as means to advance fairness and justice (Goodwin, 2005). According to Goodwin, there are three reasons for admitting all candidates to a weighted lottery:

1. For every candidate submitted to the test there is *some* chance that their score does not reflect their true merit, either because of marking error, or because the test is not well-designed. Specifically, for candidates scoring just outside Boyle's "border zone", there is a good chance that their true merit is very close to those who scored just within the "border zone", and therefore they should be admitted to the lottery as well. This argument can be repeated until all candidates are admitted to the lottery.

2. From certain justice perspectives, no one should be barred from success *ab initio* due to lack of talent.[14] In a weighted lottery, no matter how bad your chances are, you have at least some chance of winning.

3. If, as Boyle argues, it is useful to be aware of the chance element in testing and selection, would not all candidates, rather than just the borderline candidates, benefit from this awareness? The beneficial effect of restricting the pride of winners and the despondency of losers should be applied to all.

[14]This is not true for all perspectives of justice. More about how lotteries fit with various perspectives of justice and fairness is available in Goodwin (2005); Saunders (2008).

Goodwin's criticism focuses entirely on issues of fairness and justice. This makes sense in the context of an education system, as education is often considered a mechanism for advancing social justice and fairness, e.g. in providing equal opportunities. The applicability of such arguments to the science funding case is more limited, and the scope of this thesis excludes the question of the role of science funding in promoting social justice. For purely pragmatic reasons a restricted lottery in a border zone seems more efficient, especially if the border zone is small enough to be treated with a simple (equal chance) lottery instead of a graduated lottery. However, experience with the system in practice will provide further insight into the differences between a border-zone lottery and a full lottery, and this thesis does not reject the viability of a full lottery as a potential allocation mechanism. After all, if it can be shown that the cohorts selected by a full lottery perform no worse than cohorts selected by peer-review or border-zone lottery, then the cost-saving and fairness advantages of a full lottery will tip the balance in its favour.

Boyle's discussion of a lottery mechanism highlighted important issues in the statistical evaluation of individual ability, and in the interplay of fairness and efficiency in a mechanism that selects from a pool of individuals. The next section discuses different aspects of lotteries, from the perspective of economic theory.

6.2.2 The economics of distributing goods by a lottery

Boyce (1994) challenges the notion that when lotteries are chosen in real-world scenarios over other distribution mechanisms it is because of their fairness. He claims that in many real life situations many members of the community are excluded from participating in a given lottery, and furthermore a discriminatory fee is often required to participate in the lottery. These conditions, he argues, undermine many lotteries' claim to fairness. However, he argues, agents have reasons to prefer lotteries over other distribution mechanisms for purely self-interested reasons. His argument presents a mathematical formalism of distribution by lottery, which is compared to three other candidate distribution mechanisms: auctions, queues, and measurements of merit. As will be shown below, allocation by peer review bears some similarities both to distribution by auction and to distribution according to measurements of merit.

Before going into the details of Boyce's argument, it is important to note its relevance and limitations to the present case. Boyce develops an economic theory of the distribution of (in principle) monetisable goods, such as hunting permits. Furthermore, these goods are destined for consumption by individuals, rather than utilisation for the greater good, and the analysis of value or utility in Boyce's framework is only done from the point of view of individual consumers. These characteristics already set out a clear difference between Boyce's framework and the issue of selecting individuals to carry out research on behalf of a community. Because of these differences, the presentation below is structured such that Boyce's discussion of each distribution mechanism is followed by an analysis of an analogous mechanism in science funding, and what the key differences would be between the individual-consumption case and the social-utilisation case. While the differences make these analogies seem rather weak at times, going through Boyce's arguments will highlight important aspects of distribution mechanisms that have not been discussed until now.

Optimal distribution

First, Boyce establishes the condition for optimal distribution. Assume we have k homogeneous goods to be distributed among N people. These people will place some value on the goods, which could then be ordered to give a ranking of utilities, say from v_1 for the highest value to v_N for the lowest. In the most efficient allocation, the goods will go to those who value them the most, yielding an overall utility of $\sum_{i=1}^{k} v_i$. Boyce notes, however, that the satisfaction of those members of the group who receive the goods is only one aspect of the efficiency of a distribution mechanism. The other aspect, according to Boyce, is in communal rebate. If the k goods are provided from some collective pool, it may be preferable to require payment from the members who received the goods. This payment could then be distributed back to the community. Examples of this are discussed in the presentation of specific distribution mechanisms below.

Boyce's analysis relies heavily on the value individuals place on the good (in our case, the research grant). This is not the case in science funding, where the measure of a good distribution is one that maximises contribution to well-being via the products of research, not one that maximises the satisfaction of the desire of scientists for grant moneys. Keeping this clear distinction in mind, it is worthwhile to consider the issue of consumption in the science funding case for two reasons:

- We may consider whether there is any correlation between the consumption utility of a research grant for a particular scientist, and the likelihood of that scientist's project resulting in a significant contribution to well-being.
- If two funding mechanisms are equally good at generating contributions to well-being, we may prefer the mechanism that better satisfies the desires of participating scientists, assuming other secondary desiderata, such as fairness, being equal.

Distribution by auction

The go-to economic mechanism for the distribution of goods is an auction. As a well-studied distribution mechanism, auctions serve as a good benchmark for other distribution mechanisms, such as lotteries. According to Boyce, in a k price auction of k homogeneous goods, the goods will sell for some market value v_k. There will be k people who are willing to pay this market price, because they value the goods more than the market price, $v_i \geq v_k$; label these people group A. Each member of group A has a consumer benefit of $v_i - v_k$, leading to a total benefit of $(\sum_{i=1}^{k} v_i) - kv_k$, while the other members of the population, the ones who value the good less than its market value, have no benefit. However, the auction's earnings could be rebated to the community, in which case, assuming equal rebate, there will be a further individual benefit for all members (including members of group A) equal to kv_k/N. Note that for large communities ($N \gg k$) this benefit vanishes.

An analogous system to an auction in the science funding case would be if scientists had to make certain promises about future utilisation of the funds in order to win them, the grants going to those scientists who promised the most. In this case, the scientists would "pay" for the grants with their time and labour, and this "payment" will be distributed to society via the impact of their research. The sceptical results from the previous chapters significantly undermine the viability of this option, because the "payment" offered by scientists cannot be predicted or evaluated accurately in advance.

In such a "promise competition" there would be a clear incentive to exaggerate what one can deliver, with clear harmful consequences. In fact, since proposals in peer review are evaluated as a hybrid of researcher credentials, project details, and expected impact (see Chapter 1), some element of auction (in the form of promise competition), and motivation for exaggerated promise, already exists in the current peer review system.[15]

A good measure against exaggeration would be to penalise scientists who did not deliver on their promises. However, due to the highly uncertain nature of research such penalisation is likely to be dished out to scientists who gave their honest best estimate. Furthermore, penalisation could, in the long run, result in more risk-averse proposals, to the detriment of the entire enterprise (see results of simulation in Chapter 5 and Gillies' discussion of highly innovative research in §1.6). The similarities between Boyce's formal treatment of distribution by auction and the current practices of science funding offer a new, and worrying, perspective on the operation of grant peer review, and offer further motivation for exploring alternative allocation mechanisms.[16]

Distribution by queue or by evaluation of an individual's earned merit

According to Boyce, in a queue or merit system, the k individuals who value the goods the most will need to spend resources by an amount close to v_k in order to win the goods. The kind of queue discussed here is a first-comes-first-served mechanism, where individuals can spend resources (waking up earlier, sleeping by the venue the night before) to improve their chances of winning the goods. From an economics perspective, this mechanism's operation is indistinguishable from a merit evaluation system, if we assume a merit system where the individuals are able to expend resources in order to gain merit. The case where merit is not obtainable by all is discussed in the next section. A queue has a similar individual efficiency performance as an auction, because k individuals win the goods by "giving up" v_k worth of resources. However, queues are less efficient from a community perspective, since the cost paid by participants is dissipated (lost) in the case of queues, without leaving the possibility of communal rebate.

Boyce only considers the merit evaluation mechanism in terms of a) its ability to assign goods to those how value them the most and b) its cost of operation. A first-comes-first-served basis for science funding will indeed be dominated by these factors, but would completely neglect the important issue of potential utilisation by the winning scientists. Despite the worries regarding prediction of impact presented in earlier chapters, even the modelling work of the previous chapter admitted that some gains can be made by considering the merit of proposals, i.e. the expected utilisation of research funds.

A science funding analogy of merit evaluation is, of course, peer review, though it is not clear whether we are talking about earned merit or natural merit (see discussion in the next section). Merit evaluation in science funding is clearly more important than the role assigned to merit evaluation by Boyce, because of the importance of utilisation which is ignored in Boyce's analysis.

In addition, in the science funding case there may be a further consideration, which is the advantage, both to applicants and reviewers, of participating in the review process. For the applicants, these benefits include constructive criticism from experts in their fields who they

[15] The issue of exaggerated promises by scientists and the harm caused by the resulting unrealistic expectations is discussed in several of the papers collected by Irwin and Wynne (1996).

[16] An alternative auction-like mechanism, where scientists compete by proposing sensible cost-saving mechanisms in order to win grants, would possibly help as a one-off exercise to curtail inflating expenses such as instrumentation costs. However, it is not likely to be a sustainable allocation mechanism.

might not have access to otherwise, and, arguably, a more honest opinion of their proposal allowed by the anonymity of the review process. As to reviewers, the process grants them access to a comprehensive snapshot of the research agenda in their field, which is fuller than the picture derived from the list of accepted proposals (which is often made public), and timelier than the published record due to the duration of research and delays in the publication process itself. Furthermore, being a member of a review panel grants the reviewers prestige as experts in their field, and provides them with tacit knowledge about the workings of the system which might help the chances of their own proposals or those of their colleagues.

Having said that, it is not clear that these advantages are significant when compared to issues of utilisation and cost, or even desirable, nor is it clear that these benefits cannot be captured in other distribution systems, or via pathways outside the distribution mechanism. The advantages to reviewers, for example, are also present in the triage system proposed in this chapter.

Distribution by evaluation of an individual's natural merit

According to Boyce, from an economics perspective there is no difference between distribution by inherent merit and distribution by lottery, except that individuals do not have equal chances to win the good. By this he means that k individuals will win the k goods based on a parameter (inherent merit or the lottery result) which is independent of the measure of efficiency, namely the utility that individuals assign to the goods. The economics of distribution by lottery are discussed in greater detail below.

As Boyce notes, there is a distinct difference between merit allocation when all levels of merit can be obtained by all individuals for different costs (earned merit), and when some (high) levels of merit are exclusively attainable by certain individuals (natural merit). In the case of science it seems clear that *some* natural ability is required, but present gatekeeping mechanisms, e.g. requiring a PhD in the topic, already test for the necessary level of ability, such that it does not differentiate further between the target population. Few ($l \ll k$) individuals may sometimes rank far higher than the rest of the cohort in terms of natural ability, but in such cases it might be possible to identify them in an early stage, e.g. via international youth competitions, known open problems in the discipline, or recognition by teachers, and allocate to them the first l grants, leaving the remaining $(k - l)$ to be allocated by whatever distribution mechanism is chosen.

The topic of exceptional talent and skill, and researcher ability in general, has been left out of the models presented in the previous chapters for simplicity's sake. However, following the arguments in Boyle's and Boyce's works, its importance is clear. The topic is addressed more fully as part of the proposal presented in §6.3.

Distribution by lottery

First, Boyce establishes that lotteries are not efficient, in the sense that they do not maximise overall utility. For now, assume the lottery is non-transferable, i.e. winners cannot sell their winnings to other members of the community. The overall utility yield will be the average utility multiplied by the number of goods, $kE(v)$. It is easy to see that this quantity is always smaller or equal to the optimal utility presented above, and it is only equal when everyone values the goods the same.

Boyce then extends his analysis to a consideration of community rebate in the lottery case. If the lottery requires that participants pay a fixed, non-refundable fee F, the number of participants

in the lottery, n, will be determined such that the last person to participate is indifferent between the expected value of the lottery and the fee, $F = (k/n)v_n$. All participants other than the last have positive expected utility, as $v_i \geq v_n$ for all $i < n$. Define group B as those $n - k$ individuals who would participate in a lottery, but would not pay the market price in an auction of the same goods (note that their number, but not their identities, is the same as those who participate in the lottery and lose). For everyone in group B, $v_k > v_i \geq v_n$. Thus, if the fee was set equal to v_k, the lottery would become equivalent to an auction. Like an auction, a lottery can also implement a rebate, where the earnings from the fees are redistributed back to the community. In the absence of rebate, all members of group B would prefer a lottery to an auction, as it gives them a positive expected utility.

Now consider the case of a transferable lottery, where winners are allowed to sell their winnings to another member of the community. All community members outside of group A will, upon winning the lottery, end up selling their winning to a member of group A. Thus, a transferable lottery encourages speculating, and the number of participants in a transferable lottery will be greater than the number of participants in a non-transferable lottery.

First, let us consider the issue of transferability in the science case. In Boyce's analysis the goods are non-monetary and the agents obtain them with money, whereas in the science funding case the goods are composed of a significant monetary element (as well as some non-monetary perks) and the scientists obtain them by writing proposals, a process which dissipates their time. Collaborations aside, scientists do not seem particularly interested in obtaining each other's time, making transferability problematic. I will therefore consider only non-transferable lotteries as possible science funding mechanisms.

Now, consider the possibility of participation fees and community rebate in the science funding case. Currently, research proposals have little value for anyone except, perhaps, their author, and so there is no possibility of rebate (as is common in merit evaluation systems). In order to consider possible rebate mechanisms, the time spent competing for grants needs to be used to achieve something of value to the community (see examples below).

A lottery proposal for science funding

As a summary of the lessons from Boyce's analysis of costs and communal rebate, let us consider what a lottery proposal for science funding might look like. A more detailed proposal that takes into account the other perspectives discussed in this thesis is presented later in this chapter. Based on Boyce's paper, and remembering the issue of utilisation absent from his account, we may propose admitting into a science funding lottery all those individuals who:

1. Hold relevant qualifications, e.g. a PhD or postdoc in the topic, relevant publications, or a university position.

2. Are able to produce a coherent research plan.

3. Are willing to complete a common task set to all applicants, which serves the community interests, e.g. contributing their time and experience to the education system or relevant industries, or mentoring young researchers.

The first two items on the list serve as filters for natural and (some) acquired ability, and bear resemblance to Boyce's merit evaluation distribution mechanism, i.e. they dissipate the time

spent by applicants with no direct communal rebate. These checking mechanisms are very similar to the checking mechanisms proposed by Gillies (2014) for his random allocation mechanism (see §1.6). These (light) evaluation mechanisms are required to promote effective utilisation of the funds, and they still offer significant cost-saving over current peer review practices by requiring much less information, as they drop the requirement for the extra information necessary to decide *between* proposals. The production of a coherent research plan would also help identify outstanding candidates which should be exempt from the lottery and awarded the grants directly, perhaps in mixture with "scouting" processes outside of the funding allocation mechanism (more on this point later, as it stands in tension with some of the work presented in earlier chapters). The third item on the list reflects a non-monetary equivalent of the lottery participation fee, implementing the equivalent of a societal rebate.

Such a system would be preferable to the current system of peer review, cost-wise, if the total amount of time spent producing the above is lower or equal to the amount of time currently required to write grant proposals.

The two sketches of lottery-based science funding mechanisms, presented above in response to Boyle's and Boyce's papers, are combined and developed into a mechanism proposal in the next section.

6.3 Design of a possible science lottery mechanism

The previous section presented two theoretical approaches to the use of lotteries, and each could be, with some modification, applied to the case of science funding. Another important lesson from the works presented in the previous section is the importance of small details that can make a big difference between two setups that could both be called "lotteries". Given the strong motivation to explore lottery elements in a science funding mechanism, based on the results of the previous chapter, this section presents a sketch of one possible design of a lottery mechanism for distributing research grants; this sketch is made in order to highlight the various considerations that are involved, and to show how the results of previous chapters can bear on the design of a science funding mechanism.

6.3.1 Organise panels by epistemic activities

Selection of applicants depends on the skill set required of the applicant, and on the similarity of the proposed project to previously attempted projects. Both of these judgements, of required skills and of similarity to past projects, require knowledge of a specific area of science. Thus, it makes sense to have the funding mechanism operated by multiple sub-organisations, each responsible for a specific area of research, in a similar manner to the different funding panels within the NSF. However, due to the dynamic nature of research, this structure should be subjected to constant revision, as new areas emerge and old areas diminish in significance.

Based on the expert knowledge required, it makes sense to assign panels according to different epistemic activities (Chang, 2012), i.e. based on the activities the researcher is likely to perform in the practice of their research, rather than, say, academic disciplines or addressed social need, which may group together researchers who engage in a heterogenous set of practices. Communities engaged in a particular epistemic activity are best positioned to accumulate and access knowledge regarding the relevant skill set and similar past projects, and are therefore best positioned, within

the research community, to evaluate proposals. Examples of epistemic activities in this context include the design of computational models of climate systems, the construction of optical tools (such as optical tweezers) for the study of biological and chemical colloids, and the observation of particular species in their natural habitats. In this, I accept some aspects of Polanyi's arguments regarding science funding (presented in §1.5), stemming from the role of tacit knowledge in epistemic activities, though in general the mechanism proposed here significantly differs from the peer review he defends, as discussed below.

6.3.2 Initial filter by fair and public criteria

Scientific activity is highly specialised. As such, most members of society would not make good utilisation of science grants. Luckily, scientific activity, and especially scientific training, is also highly codified, in university courses, postdoc programs, and counting of publications and citations. While each of these codified practices has limitations as a measure of ability, combinations of indicators could offer a range of tools for individual panels to create fair and public criteria required to submit a funding application. For example, some panels may require a PhD from a set of recognised institutes, others may add a requirement for a certain number of publications in a set of relevant journals, etc. When drafting these requirements, it is important that elements of chance and bias (e.g. in getting a publication) are remembered, and to the extent that this is possible multiple alternative routes are offered for candidates to meet the criteria. Furthermore, the discussions about requirements should take place openly and frequently within the active community pursuing the system of practice, and should preferably focus on the *minimal* set of evidence that can guarantee the applicant has the *minimal* skill set required to pursue research in the area.

There are two main reasons for focusing on the *minimal* set of skills, as opposed to a *desired* set of skills or an *evaluation* of skill to go along the evaluation of the proposal:

1. All else being equal, a broader admission into the system will increase fairness and representation, and will increase the likelihood of the lottery admitting unorthodox individuals with unorthodox ideas.

2. Given current tools and understanding, our ability to state exactly, in advance, what the required skill set would be is limited, and our ability to measure those skills even more limited.

The second point has been discussed at length in a previous work (Avin, 2010). In brief, it is the result of the following considerations:

- Scientific activity, starting with funding and ending with publication of results, is extremely heterogeneous, requiring, among others, technical skills, cognitive skills, interpersonal skills, managements skills, emotional resilience, creativity, and discipline.

- Some of these skills are measurable, but such measurements (e.g. in the screening of candidates for high-rank positions in the Israeli army, including non-combatant positions) can be very costly, requiring a trained psychologist to spend several intensive days with the candidate while the candidate performs various tasks in special test facilities.

- Many of these skills are difficult the operationalise, as there are different views about what these skills mean and how they are manifest.

- Some skills are often latent, only made manifest in rare situations that are hard to recreate in a test environment.

- Some skills may change over time, due to personal development, personal trauma, or other sources; significantly, the change may occur *during* the length of a research project, which is often measured in years.

- The strength of some of these skills may be highly situation-dependant, relying less on the individual and more on the physical or social context of the lab, such that they should not serve as a basis for selection.[17]

- The relevance of some of these skills depends on the specific nature of the research project, but as discussed in previous chapters, there is high uncertainty about the precise nature of the project *ex ante*, at the point of proposal evaluation.

Despite all the above limitations, it is hard to argue that there are no cases of robust high ability in individual scientists. Such cases are given special consideration in the proposal, as discussed below. If further evidence suggests there really are no such cases, these special provisions may be dropped.

6.3.3 Use short proposals to locate projects on the epistemic landscape

As has been extensively argued in the previous chapters, uncertainty is inherent to scientific research. Therefore, it makes no sense, neither for accountability nor for efficiency, to ask candidates for detailed research plans. Still, not all projects are identical; history tells us that some projects yield great benefits to society and further research, others less so, as has been captured by the notion of epistemic fitness developed in earlier chapters. As a compromise, it makes sense to ask candidates for short project descriptions, that associate the project with the panel it is submitted to, that outline the perceived potential of the project, and that detail its similarity to past projects, or lack thereof.

Such proposals should serve four purposes, and no other:

- Validate that the project was assigned to the right panel, and if necessary refer it to another panel.

- Further validate that the applicant is minimally conversant in the knowledge of the field, and outright reject applications from candidates which are not. This should be done carefully however, as radically novel proposals (proposals that lie outside the "vision range" of the panel members) may appear at first incomprehensible or incompetent.

- Locate, as accurately as possible, the proposal within the best estimate of the epistemic landscape of the domain. This largely involves drawing analogies to similar past projects and their revealed fitness gains, and some extrapolation into the future of the field and the expected fitness of the proposed project. Since information provided in the proposal is slim, the assignment should be rough, into groups of "known high fitness", "known medium fitness", "known low fitness" and "unknown fitness". It might be possible to introduce

[17] This point about situation-dependent personal traits bears strong resemblance to the situationist account of moral character presented by Doris (2002).

graduation within the unknown fitness group as well, if the distinction between known and unknown is done on a scale rather than as a sharp distinction.

- Contribute to the detection of rare cases of exceptional talent or skill, where the application should be funded outright.[18] Preferably, the main bulk of the detection of exceptional skill should occur outside of the funding exercise, e.g. via international competitions, or if a talented individual successfully solves a "hard nut", a long-unsolved problem in the discipline, or if they are able to make a significant and recognisable novel contribution without guidance or financial aid. If these signs are not detected prior to the funding exercise, a research proposal may indicate either of the last two, and panel members would be allowed to inquire further into such cases.

6.3.4 Triage proposals, using a lottery for the middle group

As described in §6.2.1, the assignment of expected fitness, based on the location of the project in the epistemic landscape, is used to triage the proposals:

1. All proposals of known high fitness should be funded. Based on the results of Graves et al. (§6.1.1), this would account for about 10% of proposals, though of course some variation is expected over time and between fields.

2. Proposals of known medium fitness and proposals of unknown fitness should be placed in a lottery. If graduation is used for the unknown fitness group, a graduated lottery may be used accordingly, in a similar manner to Boyle's graduated lottery.

3. All proposals of known low fitness should be rejected. Based on Graves et al. this would account for 50-60%.

Further fine details should be considered:

- The lottery should be carried out publicly, and the random selection mechanism should be open to scrutiny.

- Authors of applications which have been scored as known low fitness should be informed of the past projects which have been relied upon to make the judgement.

- If there are not enough funds to fund all projects of known high fitness, e.g. in the early stages following a major breakthrough, it may be preferable to hold back and only select a significant portion of these proposals (by lottery). This will allow non-paradigmatic research (the unknown fitness group) a chance of funding, and will also help prevent over-specialisation of the domain. The high fitness projects which are left unfunded in that particular round are likely to be funded in near-future consecutive rounds, when more fine-grained information will be available about the epistemic landscape near the high fitness peak.

[18] Examples of cases where short texts were sufficient to detect exceptional talent include Hardy's recognition of Ramanujan, and Russell's recognition of Wittgenstein. However, the error rate for such cases can be quite high, as seen in the historical examples of poor evaluation presented by Gillies (see §1.6), and therefore selection via this process should be preferably combined with other indications of exceptional talent, and the performance of selected individuals should be monitored.

6.3.5 Managing potential outcomes of introducing a lottery

There may be initial upheaval following the introduction of explicit random selection into a hitherto (apparently) fully decision-based selection mechanism, either from scientists themselves, or from the general public and its representatives about the apparent misuse of public money. This may be counteracted by communicating the message of this thesis: uncertainty in research is ineliminable, and a limited lottery has a good chance of yielding better results for society in the long run.[19]

Two expected objections to the proposal are related to waste: one worry is about an increase in the number of low-quality proposals funded, the other worry is that a lottery may encourage malicious abuse of the system, i.e. applicants submitting off-hand proposals, winning by lottery, and then wasting the funds. First, it is important to note that even under the current system there are projects that lead nowhere, and scientists who misuse public funds. Second, both worries can be mitigated by follow-up monitoring post-funding by the funding agency, especially of projects funded by lottery, e.g. by requiring annual reports and utilising occasional spot checks of laboratories. If the will and funds could be mustered, this exercise could be extended from a mere policing effort to a continual communication and a positive supporting role the funding body could offer the researchers they fund, a role they are particularly suited for, given their connections to field experts and their knowledge of the current research portfolio.

Finally, a serious concern is that projects have high set-up costs, and that the regular freezing and unfreezing of projects that can be expected under a lottery system will be highly inefficient. This concern is somewhat lessened by the triage element, as proposals for continuation are likely to have known fitness, and therefore if that fitness is high they would be funded without a lottery, and if that fitness is not high then perhaps the loss is not so great.

6.4 When should a lottery not be used

The argument for a lottery relies on various assumptions about the nature of research. It is possible that in certain domains these assumptions do not hold, and therefore allocation of research funds by lottery will not be a good method. Such domains might be identified by the kind of projects being proposed, or by the kind of discipline in which projects are proposed. This section looks at some of these scenarios.

6.4.1 Very expensive projects

The lottery mechanism was designed with a certain project size in mind, which is the size of projects applied for in NIH or NSF grants. These projects last anywhere from one year to seven years, and cost in the range of tens of thousands of dollars to a few million dollars per year. In contrast, some science/engineering mega-projects, such as the Human Genome Project, cost much more per year and last for a much longer time. There are several reasons why it might not be beneficial to include such mega-projects in a lottery system:

1. Mega projects require sustained funding over a long period of time. It is not immediately obvious how this could be guaranteed under the lottery system. For example, if a single

[19] At least as far as philosophers of science (and the few scientists) who attend philosophy of science conferences are concerned, there seems to be no serious upheaval upon hearing the proposal, though of course the reactions to *ex cathedra* arguments may differ significantly from reactions to the real thing.

lottery win locks funding for a mega-project for its entire duration, and in a short span of time many mega-projects win the lottery, then the funding pool will be tied down to these projects, crowding out all non-mega-projects in the funding pool, and the lottery's advantages of innovation and responsiveness will be lost. If, on the other hand, mega-projects would require sequential lottery wins for sustained support, we run the risk of wasting significant funds on partial projects.[20]

2. Mega projects often combine a multitude of sub-projects, some of which are purely scientific/exploratory and many others which are purely engineering. A top down approach has been shown to produce useful results in the management of large-scale engineering projects, and so it may be more efficient to submit only the exploratory scientific sub-projects to a lottery within the general budget of the mega-project (though see discussion of bounded uncertainty below).

3. Decisions to fund mega-projects often take into consideration factors that have been largely neglected in this thesis, such as job creation, national pride and/or international cooperation, and excitement and encouragement of individuals to engage with science and scientific careers. These factors place such decisions quite visibly on the political agenda of local and national policy makers, who are in a position to make a justifiable decision on matters of relatively low-uncertainty, such as job creation (at least, in this they can outperform a lottery).

6.4.2 Bounded uncertainty

In certain cases the inherent uncertainty of research is less relevant to project choice because the range of possible projects is bounded by some external constraint. For example, the research may be focused on producing a certain tool or answering a certain question within a given (short) timeframe, e.g. research into an ongoing epidemic. In such types of research the framing of the project prevents any significant exploration of uncertainties or open-ended avenues. In such cases a lottery would not prove beneficial, except possibly as a time-saving mechanism in prioritising nearly equivalent approaches. Within the target area of activity for this thesis, that of the public support of basic research (see Chapter 1), such cases are not the norm.

6.4.3 Fully explored area

When an area of research is known to be fully explored, the epistemic landscape will be fully visible, and a lottery will be worse than direct selection of projects. In such cases, however, passive mode peer reviewed applications would also not be optimal, as the field's experts have full knowledge of which are the promising projects, and can simply assign them to the most able researchers, or allow researchers to compete for them. Note, however, that such areas are likely to be quickly exhausted, leaving behind a barren epistemic landscape. It is hard to give an example, due to the inherent fallibility of all knowledge, but close approximations would be the exploration of the properties of a specific mathematical body of interest or a specific minimal axiom system, or tweaking the design of a well known instrument such as the light microscope, or sifting for novel features of a well explored data set such as a small viral genome.

[20] Current funding practices also sometimes fail in providing sustained support for mega-projects, for example the partially-funded Superconducting Supercollider in the USA.

6.4.4 Researcher identity determines fitness

As discussed in Chapter 3, the fitness of a project is a measure of the fit between societal needs and the causal consequences of the projects' results. The causal chain that follows the completion of a project is to some extent determined by the diffusion of the information, i.e. its acceptance by the scientific community and its spread by various media. There are many cases where the success of such diffusion of the information depends on the identity of the investigator who carried out the research, i.e. their track record, charisma, connections, etc. Thus, the identity of the investigator affects the causal chain from funding allocation to research-based activity, and ultimately influences the fitness of the project's results. Following Latour (1987); Kitcher (1993); Goldman (2001), it is clear that in all areas of research the identity of the researcher has *some* bearing on the eventual fitness of the project, because the researcher's authority influences the effect the research will have have on society. Nonetheless, the hope is that this influence by authority is not the dominant factor, and the actual content of the result carries more influence on the eventual impact on society's well-being. However, it is possible that this is not the case for all fields of science.

In areas where the researcher's authority strongly determines the fitness of the results they produce, a lottery would perform worse than other selection methods, though so would a peer review system that hides the identity of the applicant.

Another way the researcher identity could determine fitness is if a rare natural ability or gained skill is required to make advances in the field, for example an anthropological study of a secluded tribe that requires years of acclimatisation from both tribe and researcher, or a psychological self-study by a high functioning individual with a rare mental abnormality. In such a case a lottery would clearly be a bad choice, unless participation in the lottery depends on having the required ability or skill.

Conclusion

The previous chapter has raised, as the result of a highly idealised model, the suggestion that a lottery may outperform other mechanisms for choosing research projects. In this chapter various aspects of implementing such a lottery have been explored, suggesting that under the right conditions a lottery might be a fairer and more efficient allocation system. Of course, the lottery should remain within bounds, and not be applied in fields that would not benefit from it.

If it is ever implemented, empirical evidence gathered about the lottery payoffs and reactions to it may prove illuminating to science studies. However, even if a lottery is never actually implemented, consideration of the proposal should remind us of the key role uncertainty plays in the dynamics of scientific research.

Conclusion

This thesis argues that public support of basic research by peer review is less than optimal. While accepting Polanyi's basic argument, that no agent is better placed to pass judgments of scientific merit than scientific peers, the work shows that in important cases random allocation would fare better than human judgement. There are two main epistemic reasons for this. The first is that judgements of project merit rely on past experience with similar scientific projects, knowledge which may be very partial given the entire scope of novelty in research. The second is that the merit of projects can change over time, and the rate of change may be faster than the duration required for successfully completing a unit of research. Thus, peer assessment of merit, which is essentially a prediction of the merit of the results of proposed projects, can be undermined by changes of merit that will take place between the initiation of a project and its successful completion.

The results mentioned above were arrived at with the help of models. The use of models was twofold throughout the thesis: existing models of relevant phenomena or of analogous structure have been explored and criticised, and I have offered revised versions or novel models in their place. The existing models that are most relevant to this thesis are models of social phenomena in science, especially of project choice, as discussed in Chapter 2. Early models of project choice were shown to rely on an erroneous "lost dog" analogy, involving problematic assumptions regarding the small number of alternative projects and our ability to know their value in advance. The work of Weisberg and Muldoon (2009) addresses these problematic assumptions by shifting to a model of a population of investigators on an epistemic landscape. However, this latter model employs a vague notion of "epistemic significance", and fails to account for the dynamic nature of project merit, and is therefore unsuitable for direct application to the problem of science funding.

Noting the promise of Weisberg and Muldoon's model for the case of science funding, I turned to an exploration of an analogous, and well-studied, set of models: fitness landscapes in population biology. Using a stepwise modelling strategy, I first developed the "information landscape" model, where coordinates are associated with different possible contents of public corpuses of information, and distances are given by the intuitive conceptual similarity between the worlds they describe. Each possible set of corpus contents is associated with a measure labelled "epistemic fitness", which measures the fit between the causal consequences of the existence of the information and the goals and values of the society which utilises the information. The method used to evaluate this "fit" will depend on the model-user's notion of well-being, and is left unspecified by the model. Using the information landscape model as an intermediary, I presented a revised version of Weisberg and Muldoon's "epistemic landscape" model. The introduction of "epistemic fitness" addresses the first deficiency of Weisberg and Muldoon's model, in defining a new "merit" measure that is (somewhat) precise and that can be related to

the aim of public science funding bodies. I showed how the revised epistemic landscape model can represent the decisions of funding bodies, which aim to select the projects of greatest fitness contribution potential.

Using the notion of epistemic fitness and the revised epistemic landscape model, Chapter 4 explored the ways in which epistemic fitness can change over time. The exploration relied on three historical examples: the invention of the laser gyroscope, the discovery of the double-helix structure of DNA, and the responses to potential health and environmental risk in the development of recombinant DNA. From these examples the analysis abstracted eleven different processes via which the fitness of projects can change over time, and these only include changes that occur as a response to research activity. The existence of so many causal processes that can alter epistemic fitness implies that the assignment of fitness is complex, and that individuals attempting to predict the future fitness of projects are likely to not have the information required to make sufficiently accurate predictions. To emphasise this point, and make it more tractable, Chapter 5 used the revised epistemic landscape model to develop a computer simulation of different funding mechanisms. The simulations showed that on large dynamic landscapes, selecting projects at random outperforms the selection of the highest fitness projects from those which are similar to past projects. The implications of this result are significant: within a set of idealising limitations, a selection of projects that relies on past experience, such as can be expected from peer review, performs worse than random selection. This is the main result of the thesis, and it provides substantial backing to the argument against peer review made by Gillies (2014).

Building on this result, the last chapter engages in an exercise of institutional design. In designing an alternative funding mechanism, I bring to the fore other aspects of science funding that have been set aside in the modelling work of earlier chapters. If a lottery is not less effective than peer review (and may in fact be more effective), other considerations, such as cost and fairness, tip the balance towards a lottery-based system. Some aspects of peer review are maintained, for initial filtering and accountability, but the sceptical result is used to make informed use of the information provided by peer review, keeping the reliance on it within bounds of its limited predictive power.

Ultimately though, it must be acknowledged that the bulk of the argument for employing a lottery for science funding relies on non-empirical arguments. It is therefore suggested that further empirical work is done on this issue, either by implementing a small-scale lottery fund, by conducting laboratory tests simulating the conditions of peer review panels, or by detailed historical or field studies of this important aspect of contemporary scientific life. The philosophical investigation of the public funding of basic research presented in this thesis provides ample motivation and justification for carrying out such empirical studies.

Appendix: Source code for simulations

```
from numpy import *
from numpy.random import *
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
import random

# return a 2d matrix with a peak centred at loc, with width sig, of max_height
def get_peak(size, loc, sig, max_height):
    m = mgrid[:size,:size]
    biv = mlab.bivariate_normal(m[0], m[1], sig[0], sig[1], loc[0], loc[1])
    return biv*float(max_height)/biv.max()

class Landscape(object):
    def __init__(self, size):
        self.size = size
        self.matrix = zeros((size,size))
        self.vision = zeros((size,size))

    def get_individuals(self, num):
        return hstack( (randint(self.size-1, size=(num,2)), # starting positions
                        randint(max(1,self.avg_countdown/2), # countdowns
                        int(self.avg_countdown*1.5)+1,
                        size=(num, 1))) )

    def set_individual_vision(self, ind):
        x = ind[0]
        y = ind[1]
        for xi in range(-1,2):
            for yi in range(-1,2):
                if (x+xi < self.size) and (y+yi < self.size) \
                    and (x+xi >= 0) and (y+yi >= 0):
                    self.vision[[x+xi],[y+yi]] = 1

    def init_individuals(self,num, avg_countdown=3):
        self.avg_countdown = avg_countdown
        self.individuals = self.get_individuals(num)
        self.accumulated_fitness = 0
        for ind in self.individuals:
            self.set_individual_vision(ind)
```

```
def print_matrix(self):
    print self.matrix

def plot(self,individuals=False):
    m = mgrid[:self.size,:self.size]
    plt.contourf(m[0],m[1],self.matrix,range(self.size*2),cmap=plt.cm.gray)
    if individuals:
        for ind in self.individuals:
            plt.text(ind[0],ind[1],str(ind[2]))

def show(self, individuals=False):
    self.plot(individuals=individuals)
    plt.colorbar()
    plt.show()

def save(self, filename, individuals=False):
    self.plot(individuals=individuals)
    plt.colorbar()
    plt.savefig(filename, bbox_inches=0)
    plt.show()

def plot_vision(self):
    m = mgrid[:self.size,:self.size]
    plt.pcolormesh(m[0],m[1],self.vision,cmap=plt.cm.hot)

def show_vision(self):
    self.plot_vision()
    plt.show()

def countdown_individual(self,ind):
    ind[2] -= 1 # reduce countdown
    if ind[2]: #countdown not zero
        return ind, -1 # new_z as -1 indicates the individual hasn't moved

    # set new countdown
    new_countdown = randint(max(1,self.avg_countdown/2),int(self.avg_countdown
        *1.5)+1)
    ind[2] = new_countdown

    ind_z = self.matrix[ind[0],ind[1]]

    # contribute to accumulated fitness based on current location
    self.accumulated_fitness += ind_z

    return ind, ind_z

def move_individual(self,ind):

    end_x = start_x = ind[0]
    end_y = start_y = ind[1]

    max_z = self.matrix[start_x,start_y]

    # move to highest neighbour (including current position)
```

```
        for x_offset in range(-1,2):
            cur_x = min(self.size-1,max(0,start_x+x_offset))
            for y_offset in range(-1,2):
                cur_y = min(self.size-1,max(0,start_y+y_offset))
                if self.matrix[cur_x,cur_y] > max_z:
                    max_z = self.matrix[cur_x,cur_y]
                    end_x = cur_x
                    end_y = cur_y
        ind[0] = end_x
        ind[1] = end_y

        self.set_individual_vision(ind)

        return ind

    def step(self):
        for i in range(len(self.individuals)):
            self.individuals[i], ind_z = self.countdown_individual(self.
                individuals[i])
            # If individual transcends a certain height, slightly disrupt the
                landscape
            if ind_z > 1:
                self.matrix = self.matrix+ random_integers(-1,2,(self.size,self.
                    size))

            if ind_z != -1:
                self.individuals[i] = self.move_individual(self.individuals[i])

class GaussianLandscape(Landscape):
    def __init__(self,size,num_peaks,max_height):
        super(GaussianLandscape,self).__init__(size)
        m = mgrid[:self.size,:self.size]
        peaks = randint(1, self.size-1, size=(num_peaks,4))
        self.matrix = zeros((size,size))
        for peak in peaks:
            self.matrix += get_peak(self.size, (peak[0], peak[1]), (peak[2],peak
                [3]), random.random())
        self.matrix *= max_height / self.matrix.max()

    def add_gaussian(self, loc, sig, height): #height can be negative
        self.matrix += get_peak(self.size, loc, (sig,sig), height)
        # do not allow negative values
        self.matrix = self.matrix.clip(min=0)

    def winner_takes_all(self, ind):
        loc = [ind[0],ind[1]]
        height = self.matrix[loc[0],loc[1]] * -1 # lower current position by its
            height
        sig = self.size*0.01
        self.add_gaussian(loc, sig, height)

    def reduced_novelty(self, ind):
        loc = [ind[0], ind[1]]
        height = self.matrix[loc[0],loc[1]] * -0.5 # lower local neighbourhood by
```

```
        majority of height
    sig = self.size*0.05
    self.add_gaussian(loc, sig, height)

def new_avenue(self):
    loc = randint(self.size-1, size=(2))
    height = randint(self.matrix.max())
    sig = self.size * 0.06
    self.add_gaussian(loc, sig, height)

def step(self, cutoff=0.7, funding='best', dynamic=True):
    individual_indexes_to_move = []
    for i in range(len(self.individuals)):
        # update individual's countdown, check if research finished
        self.individuals[i], ind_z = self.countdown_individual(self.
            individuals[i])

        if ind_z != -1:
            # Effects triggered above cutoff
            if float(ind_z)/self.matrix.max() > cutoff:
                if dynamic:
                    self.reduced_novelty(self.individuals[i])
                    self.new_avenue()
                else:
                    pass

            # Always triggered effects
            if dynamic:
                self.winner_takes_all(self.individuals[i])
            else:
                pass


        if ind_z != -1:
            individual_indexes_to_move.append(i)

    # move individual on new landscape
    for i in individual_indexes_to_move:
        self.individuals[i] = self.move_individual(self.individuals[i])

    ## Funding stage ##
    if not individual_indexes_to_move: # if there are no candidates there is
        no free cash
        return

    num_total_individuals = len(self.individuals)

    # remove moved individuals from current individuals, add them to candidate
         list
    old_candidates = []
    for i in individual_indexes_to_move:
        old_candidates.append(self.individuals[i])
    self.individuals = delete(self.individuals,individual_indexes_to_move,0)
```

```
# add new candidates until number of applicants == total num of
    individuals
new_individuals = self.get_individuals(num_total_individuals - len(
    individual_indexes_to_move))
candidates = vstack( (array(old_candidates), new_individuals) )

if funding == 'best':
    zs = [self.matrix[ind[0],ind[1]] for ind in candidates]
    zs.sort()
    zs.reverse()
    zs = zs[:len(individual_indexes_to_move)]

    count = 0
    for ind in candidates:
        if self.matrix[ind[0],ind[1]] in zs:
            self.individuals = vstack( (self.individuals, ind) )
            self.set_individual_vision(ind)
            count += 1
            if count >= len(individual_indexes_to_move):
                break
elif funding == 'best_visible':
    non_visible_candidate_indexes = []
    for i in range(len(candidates)):
        ind = candidates[i]
        if not self.vision[ind[0],ind[1]]:
            non_visible_candidate_indexes.append(i)
    candidates = delete(candidates,non_visible_candidate_indexes,0)
    zs = [self.matrix[ind[0],ind[1]] for ind in candidates]
    zs.sort()
    zs.reverse()
    zs = zs[:len(individual_indexes_to_move)]

    count = 0
    for ind in candidates:
        if self.matrix[ind[0],ind[1]] in zs:
            self.individuals = vstack( (self.individuals, ind) )
            self.set_individual_vision(ind)
            count += 1
            if count >= len(individual_indexes_to_move):
                break
elif funding == 'lotto':
    shuffle(candidates)
    candidates = candidates[:len(individual_indexes_to_move)]
    for ind in candidates:
        self.set_individual_vision(ind)
    self.individuals = vstack( (self.individuals, candidates) )
elif funding == 'triage':
    non_visible_candidate_indexes = []
    for i in range(len(candidates)):
        ind = candidates[i]
        if not self.vision[ind[0],ind[1]]:
            non_visible_candidate_indexes.append(i)

    # Use lotto on half the candidates
```

```
            non_visible_candidates = candidates[non_visible_candidate_indexes]
            shuffle(non_visible_candidates)
            non_visible_candidates = non_visible_candidates[:len(
                individual_indexes_to_move)/2]
            for ind in non_visible_candidates:
                self.set_individual_vision(ind)
            self.individuals = vstack( (self.individuals, non_visible_candidates)
                )

            num_remaining = len(individual_indexes_to_move) - len(
                non_visible_candidates)

            # Use best_visible on the remainder
            visible_candidates = delete(candidates,non_visible_candidate_indexes
                ,0)
            zs = [self.matrix[ind[0],ind[1]] for ind in visible_candidates]
            zs.sort()
            zs.reverse()
            zs = zs[:num_remaining]
            count = 0
            for ind in candidates:
                if self.matrix[ind[0],ind[1]] in zs:
                    self.individuals = vstack( (self.individuals, ind) )
                    self.set_individual_vision(ind)
                    count += 1
                    if count >= num_remaining:
                        break
        elif funding == 'oldboys':
            self.individuals = vstack( (self.individuals, array(old_candidates)) )
        else:
            raise KeyError('Unknown funding option %s'%str(funding))

        assert(len(self.individuals)==num_total_individuals)

def show_landscapes(landscapes,individuals=False):
    fig = plt.figure()
    for i in range(len(landscapes)):
        plt.subplot(len(landscapes)*100+10+i)
        landscapes[i].plot(individuals=individuals)
    plt.show()

if __name__ == '__main__':
    import sys
    option = sys.argv[1]

    size = 500
    num_steps = 50
    avg_countdown = 5
    num_runs = 5
    dynamic=True

    if option == 'show-landscape':
        size = 100
        num_peaks = 50
```

```
    height = size*2
    show_individuals = True
    num_individuals = 33
    avg_countdown = 5
    l = GaussianLandscape(size, num_peaks, height)
    l.init_individuals(num_individuals,avg_countdown=avg_countdown)
    l.show(individuals=show_individuals)
    sys.exit(0)


if option == 'compare':
    funding_options = ['best','best_visible','lotto','triage','oldboys']
    all_lines = []
    for run in range(num_runs):
        fitness_lines = []
        for funding in funding_options:
            landscape1 = GaussianLandscape(size,size/2,(size-1)*2)
            landscape1.init_individuals(int(size**0.75), avg_countdown=
                avg_countdown)
            fitness_steps = []
            for i in range(num_steps):
                if i%10==0:
                    print funding, i
                landscape1.step(funding=funding, dynamic=dynamic)
                fitness_steps.append(landscape1.accumulated_fitness)
            fitness_lines.append([funding,fitness_steps])
        all_lines.append(fitness_lines)
    if num_runs != 1:
        #average
        fitness_lines = []
        for f_i in range(len(funding_options)):
            fitness_line = []
            for step in range(num_steps):
                fitness_line.append(sum([l[f_i][1][step] for l in all_lines])/
                    num_runs)
            fitness_lines.append([funding_options[f_i],fitness_line])


    r = range(num_steps)
    plt.plot(r,fitness_lines[0][1],r,fitness_lines[1][1],
             r,fitness_lines[2][1],r,fitness_lines[3][1],
             r,fitness_lines[4][1])
    plt.legend([l[0] for l in fitness_lines],'upper left')
    plt.xlabel('simulation steps')
    plt.ylabel('global fitness')
    plt.title('Size: %d, Steps: %d, Avg. Countdown: %d'%(size,num_steps,
        avg_countdown))
    plt.show()
    sys.exit(0)


funding = option

landscape1 = GaussianLandscape(size,size/2,(size-1)*2)
```

```
landscape1.init_individuals(int(size**0.75), avg_countdown=avg_countdown)

fitness_steps = []

for i in range(num_steps):
    landscape1.step(funding=funding, dynamic=dynamic)
    fitness_steps.append(landscape1.accumulated_fitness)

plt.plot(fitness_steps)
plt.xlabel('simulation steps')
plt.ylabel('global fitness')
plt.title('Size: %d, Steps: %d, Avg. Countdown: %d'%(size,num_steps,
    avg_countdown))
plt.show()
```

Bibliography

Adriaans, P., 2013. Information. In: Zalta, E. N. (Ed.), *Stanford Encyclopedia of Philosophy*, Fall 2013 Edition.

Agar, J., 2012. *Science in the 20th century and beyond*. Polity.

Alexandrova, A., 2008. Making models count. *Philosophy of science* 75 (3), 383–404.

Alexandrova, A., Northcott, R., 2009. Progress in economics: Lessons from the spectrum auctions. In: Kincaid, H., Ross, D. (Eds.), *Oxford handbook of philosophy of economics*. Oxford handbooks online, Ch. 11, pp. 306–337.

Allen, G. E., 1975. *Life science in the twentieth century*. History of science. Wiley, New York.

Ariew, A., Lewontin, R. C., 2004. The confusions of fitness. *The British journal for the philosophy of science* 55 (2), 347–363.

Arrow, K., 1962. Economic welfare and the allocation of resources for invention. In: *The rate and direction of inventive activity: Economic and social factors*. Nber, pp. 609–626.

Arrow, K. J., 1963. *Social choice and individual values*, 2nd Edition. Wiley, New York.

Atkinson, P., Coffey, A., Delamont, S., Lofland, J., Lofland, L. (Eds.), 2001. *Handbook of ethnography*. SAGE, London.

Avin, S., 2010. Wellcome Trust Investigator Awards: Funding genius or creating genius? Part III Dissertation, University of Cambridge, Cambridge.

Bernal, J. D., 1939. *The social function of science*. M.I.T. Press, Cambridge, MA.

Bird, A., 2000. *Thomas Kuhn*. Philosophy now. Princeton University Press, Princeton, N.J.

Boyce, J. R., 1994. Allocation of goods by lottery. *Economic inquiry* 32 (3), 457–476.

Boyle, C., 1998. Organizations selecting people: how the process could be made fairer by the appropriate use of lotteries. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47 (2), 291–321.

Boyle, C., 2010. *Lotteries for education*. Imprint Academic.

Boyle, C., 2013. Examples where randomisation is being used to distribute prizes. `http://www.conallboyle.com/ExsCurrent.html`, Accessed 23 October 2013.

Brennan, P., 2013. Why I Study Duck Genitalia. http://www.slate.com/articles/health_and_science/science/2013/04/duck_penis_controversy_nsf_is_right_to_fund_basic_research_that_conservatives.html, Accessed 24 August 2014.

Brock, W., Durlauf, S., 1999. A formal model of theory choice in science. *Economic theory* 14 (1), 113–130.

Bunge, M., 1974. *Technology as applied science*. Springer.

Burch-Brown, J., 2012. Consequences, action guidance and ignorance. Ph.D. thesis, University of Cambridge.

Burch-Brown, J. M., 2014. Clues for consequentialists. *Utilitas* 26 (1), 105–119.

Burrough, P., 1989. Matching spatial databases and quantitative models in land resource assessment. *Soil use and management* 5 (1), 3–8.

Bush, V., 1945. *Science, the endless frontier: A report to the President*. U.S. Government printing office, Washington.

Calcott, B., 2008. Assessing the fitness landscape revolution. *Biology and philosophy* 23 (5), 639–657.

Carpenter, R., 1983. Scoring to provide risk-related primary health care: evaluation and up-dating during use. *Journal of the Royal Statistical Society. Series A (General)*, 1–32.

Cartwright, N., 1983. *How the laws of physics lie*. Oxford University Press, Oxford.

Chang, H., 2012. *Is water H_2O?: Evidence, pluralism and realism*. Springer, New York.

Chubin, D., Hackett, E., 1990. *Peerless science: Peer review and US science policy*. State University of New York Press, Albany.

Chubin, D. E., 1994. Grants peer review in theory and practice. *Evaluation review* 18 (1), 20–30.

Cicchetti, D. V., 1991. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and brain sciences* 14 (01), 119–135.

Cole, S., Cole, J. R., Rubin, L., 1977. *Peer review and the support of science*. WH Freeman.

Cole, S., Cole, J. R., Simon, G. A., 1981. Chance and consensus in peer review. *Science* 214 (4523), 881–886.

Collins, H. M., 2010. *Tacit and explicit knowledge*. The University of Chicago Press, Chicago.

Cormen, T. H., 2001. *Introduction to algorithms*. Algorithms. M.I.T. Press, Cambridge, MA.

Crisp, R., 2013. Well-being. In: Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*, Summer 2013 Edition.

Dasgupta, P., David, P., 1994. Toward a new economics of science. *Research policy* 23 (5), 487–521.

Davidson, D., 1980. *Essays on actions and events*. Clarendon Press, Oxford.

Davison, A. C., Hinkley, D. V., 1997. *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

Demicheli, V., Di Pietrantonj, C., 2007. Peer review for improving the quality of grant applications. *Cochrane database of systematic reviews* 2.

Diamond, A., 1996. The economics of science. *Knowledge, technology & policy* 9 (2), 6–49.

Dinges, M., 2005. *The Austrian Science Fund: Ex post evaluation and performance of FWF funded research projects*. Institute of Technology and Regional Policy, Vienna.

Doris, J., 2002. *Lack of character: Personality and moral behavior*. Cambridge University Press, Cambridge.

Duxbury, N., 1999. *Random justice: on lotteries and legal decision-making*. Clarendon Press, Oxford.

Edgeworth, F. Y., 1888. The statistics of examinations. *Journal of the Royal Statistical Society* 51 (3), 599–635.

Edgeworth, F. Y., 1890. The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53 (3), 460–475.

Elster, J., 1989. *Solomonic judgements: studies in the limitations of rationality*. Cambridge University Press, Cambridge.

Fantl, J., 2014. Knowledge how. In: Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*, Spring 2014 Edition.

Feyerabend, P., 1975. *Against method: Outline of an anarchistic theory of knowledge*. Verso Books.

Foot, P., 2002. The problem of abortion and the doctrine of the double effect. *Applied ethics: Critical concepts in philosophy* 2, 187.

Franssen, M., Lokhorst, G.-J., van de Poel, I., 2013. Philosophy of technology. In: Zalta, E. N. (Ed.), The Stanford encyclopedia of philosophy, Winter 2013 Edition. phd.

Frazier, S., 1987. *University funding: Information on the role of peer review at NSF and NIH*. US General Accounting Office.

Frigg, R., Hartmann, S., 2012. Models in science. In: Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*, Fall 2012 Edition. `http://plato.stanford.edu/archives/fall2012/entries/models-science/`.

Gallagher, M., Yuan, B., 2006. A general-purpose tunable landscape generator. *IEEE transactions on evolutionary computation* 10 (5), 590–603.

Gataker, T., Boyle, C., 2008. *The nature and uses of lotteries: a historical and theological treatise*, 2nd Edition. The luck of the draw. Imprint Academic, Exeter, UK.

Gavrilets, S., 2004. *Fitness landscapes and the origin of species*. Monographs in population biology. Princton University Press, Princeton, NJ.

Geuna, A., Salter, A. J., Steinmueller, W. E., 2003. *Science and innovation: Rethinking the rationales for funding and governance*. Edward Elgar Publishing, Northampton, MA.

Giere, R. N., 1988. *Explaining science: a cognitive approach*. University of Chicago Press, Chicago.

Gillies, D., 1992. The Fregean revolution in logic. In: Gillies, D. (Ed.), *Revolutions in mathematics*. Oxford University Press, Oxford, pp. 265–305.

Gillies, D., 2005. Hempelian and Kuhnian approaches in the philosophy of medicine: the Semmelweis case. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences* 36 (1), 159–181.

Gillies, D., 2007. Lessons from the history and philosophy of science regarding the research assessment exercise. *Royal Institute of Philosophy supplement* 61, 37–73.

Gillies, D., 2008. *How should research be organised?* College Publications, London.

Gillies, D., 2009. How should research be organised? An alternative to the UK Research Assessment Exercise. In: McHenry, L. (Ed.), *Science and the pursuit of wisdom. Studies in the philosophy of Nicholas Maxwell*. Ontos Verlag, pp. 147–168.

Gillies, D., 2012. Economics and research assessment systems. *Economic thought* 1, 23–47.

Gillies, D., 2014. Selecting applications for funding: why random choice is better than peer review. *RT. A Journal on research policy and evaluation* 2 (1).
URL `http://riviste.unimi.it/index.php/roars/article/view/3834`

Godin, B., Doré, C., 2004. Measuring the impacts of science: Beyond the economic dimension. *History and sociology of S&T statistics*.

Goldman, A. I., 1999. *Knowledge in a social world*. Clarendon Press, Oxford.

Goldman, A. I., 2001. Experts: which ones should you trust? *Philosophy and phenomenological research* 63 (1), 85–110.

Goodwin, B., 2005. *Justice by lottery*, 2nd Edition. Imprint Academic, Charlottesville, VA.

Graves, N., Barnett, A. G., Clarke, P., 2011. Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel. *BMJ* 343.

Greenberg, D. S., 1998. Chance and grants. *The Lancet* 351 (9103), 686.

Greenberg, D. S., 1999. *The politics of pure science*. University of Chicago Press, Chicago.

Greenberg, D. S., 2003. *Science, money, and politics: Political triumph and ethical erosion*. University of Chicago Press, Chicago.

Grim, P., 2009. Threshold phenomena in epistemic networks. In: *Complex adaptive systems and the threshold effect: Views from the natural and social sciences: Papers from the AAAI Fall Symposium*. pp. 53–60.

Hegselmann, R., Krause, U., 2009. Deliberative exchange, truth, and cognitive division of labour: A low-resolution modeling approach. *Episteme* 6 (02), 130–144.

Herbert, D. L., Barnett, A. G., Clarke, P., Graves, N., 2013. On the time spent preparing grant proposals: an observational study of Australian researchers. *BMJ Open* 3 (5).

Hull, D., 1988. *Science as a process*. University of Chicago Press, Chicago.

Immerman, N., 2011. Computability and complexity. In: Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*, Fall 2011 Edition. phd.

Ioannidis, J. P. A., 2011. More time for research: Fund people not projects. *Nature* 477 (7366), 529–531.
URL `http://dx.doi.org/10.1038/477529a`

Irwin, A., Wynne, B., 1996. *Misunderstanding science? The public reconstruction of science and technology*. Cambridge University Press, Cambridge.
URL `http://www.loc.gov/catdir/samples/cam034/95032980.html`

Kauffman, S. A., 1993. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, New York.

Kennedy, J. V., 2012. The sources and uses of U.S. science funding. *The New Atlantis* 36, 3–22.
URL `http://www.thenewatlantis.com/publications/the-sources-and-uses-of-us-science-funding`

Kitcher, P., 1990. The division of cognitive labor. *The journal of philosophy* 87 (1), pp. 5–22.
URL `http://www.jstor.org/stable/2026796`

Kitcher, P., 1993. *The advancement of science*. Oxford University Press, New York.

Kitcher, P., 2001. *Science, truth, and democracy*. Oxford University Press, New York.

Kitcher, P., 2011. *Science in a democratic society*. Prometheus Books, Amherst, N.Y.

Knuth, D. E., 1997. *The art of computer programming*, 3rd Edition. Addison-Wesley, Reading, MA.

Krimsky, S., 1982. *Genetic alchemy: The social history of the recombinant DNA controversy*. M.I.T. Press, Cambridge, MA.

Kuhn, T., 1996. *The structure of scientific revolutions*, 3rd Edition. University of Chicago Press, Chicago.

Laffer, A., 2004. The laffer curve: Past, present, and future. *Backgrounder* 1765.

Lakatos, I., 1970. Falsification and the methodology of scientific research programmes. In: Lakatos, I., Musgrave, A. (Eds.), *Criticism and the growth of knowledge*. Cambridge university press, Cambridge, pp. 91–196.

Lakatos, I., Musgrave, A., 1970. *Criticism and the growth of knowledge*. Cambridge University Press, Cambridge.

Latour, B., 1987. *Science in action*. Open University Press, Milton Keynes.

Lee, F. S., 2007. The Research Assessment Exercise, the state and the dominance of mainstream economics in British universities. *Cambridge journal of economics* 31 (2), 309–325.

Lenman, J., 2000. Consequentialism and cluelessness. *Philosophy & public affairs* 29 (4), 342–370.

Lipton, P., 1994. Philip Kitcher: The Advancement of Science: Science without legend, objectivity without illusions. *British journal for the philosophy of science* 45, 929–929.

Longino, H. E., 2002. *The fate of knowledge*. Princeton University Press, Princeton, N.J.

MacKenzie, D., 1998. *Knowing machines: Essays on technical change*. M.I.T. Press, Cambridge, MA.

Manson, N. C., O'Neill, O., 2007. *Rethinking informed consent in bioethics*. Cambridge University Press, Cambridge.

Martino, J. P., 1992. *Science funding: politics and porkbarrel*. Transaction Publishers, New Brunswick, NJ.

Masterman, M., 1970. The nature of a paradigm. In: Lakatos, I., Musgrave, A. (Eds.), *Criticism and the growth of knowledge*. Cambridge University Press, Cambridge, pp. 59–89.

Muldoon, R., Weisberg, M., 2011. Robustness and idealization in models of cognitive labor. *Synthese* 183 (2), 161–174.

Nagel, E., 1961. *The structure of science: Problems in the logic of scientific explanation*. Harcourt, Brace & World, New York.

Nelson, R., 1959. The simple economics of basic scientific research. *The journal of political economy* 67 (3), 297–306.

NIH, 2013a. NIH grants policy statement. `http://grants.nih.gov/grants/policy/nihgps_2013/`, Accessed 9 November 2013.

NIH, 2013b. NIH grants process overview. `http://grants.nih.gov/grants/grants_process.htm`, Accessed 18 November 2013.

NIH, 2014. Success rates - NIH research portfolio online reporting tools (RePORT). `http://report.nih.gov/success_rates/`, Accessed 11 July 2014.

NSF, 2013a. Grant Proposal Guide. `http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg`, Accessed 9 November 2013.

NSF, 2013b. US NSF - Merit Review. `http://www.nsf.gov/bfa/dias/policy/merit_review/`, Accessed 19 November 2013.

Nye, M. J., 2011. *Michael Polanyi and his generation: Origins of the social construction of science*. University of Chicago Press, Chicago.

Odenbaugh, J., Alexandrova, A., 2011. Buyer beware: Robustness analyses in economics and biology. *Biology and philosophy* 26 (5), 757–771.

Odling-Smee, F. J., Laland, K. N., Feldman, M. W., 1996. Niche construction. *American naturalist*, 641–648.

Peirce, C. S., 1879/1967. Note on the theory of the economy of research. *Operations research* 15 (4), 643–648.
URL http://www.jstor.org/stable/168276

Pigliucci, M., 2008. Is evolvability evolvable? *Nature reviews genetics* 9 (1), 75–82.

Pigliucci, M., Kaplan, J., 2006. *Making sense of evolution: the conceptual foundations of evolutionary biology*. University of Chicago Press, Chicago.

Plutynski, A., 2008. The rise and fall of the adaptive landscape? *Biology and philosophy* 23 (5), 605–623.

Polanyi, M., 1958. *Personal knowledge: towards a post-critical philosophy*. University of Chicago Press, Chicago.

Polanyi, M., 1962. The republic of science: Its political and economic theory. *Minerva* 1, 54–73.

Polanyi, M., 1967. *The tacit dimension*. Routledge & K. Paul, London.

Research Councils UK, 2006. Report of the Research Councils UK Efficiency and effectiveness of peer review project. www.rcuk.ac.uk/documents/documents/rcukprreport.pdf.

Roth, A. E., 2002. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica* 70 (4), 1341–1378.

Saunders, B., 2008. The equality of lotteries. *Philosophy* 83 (03), 359–372.

Scharnhorst, A., Börner, K., van den Besselaar, P. (Eds.), 2011. *Models of science dynamics - encounters between complexity theory and information sciences*. Springer, New York.

Sen, A., 1970. *Collective choice and social welfare*. Vol. 5. Holden-Day, San Francisco.

Simon, H. A., 1996. *The sciences of the artificial*. M.I.T. Press, Cambridge, MA.

Sober, E., 2001. The two faces of fitness. *Thinking about evolution: historical, philosophical, and political perspectives* 2, 309–321.

Stadler, B. M., Stadler, P. F., Wagner, G. P., Fontana, W., 2001. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of theoretical biology* 213 (2), 241–274.

Stadler, P., 2002. Fitness landscapes. *Biological evolution and statistical physics*, 183–204.

Sterelny, K., 1994. Science and selection. *Biology and Philosophy* 9 (1), 45–62.

Stone, P., 2011. *The luck of the draw: The role of lotteries in decision-making*. Oxford University Press, Oxford.

Strevens, M., 2003. The role of the priority rule in science. *The journal of philosophy* 100 (2), 55–79.

Strevens, M., 2011. Economic approaches to understanding scientific norms. *Episteme* 8 (2), 184–200.

Suppe, F., 1989. *The semantic conception of theories and scientific realism*. University of Illinois Press, Chicago.

Teller, P., 2001. Twilight of the perfect model model. *Erkenntnis* 55 (3), 393–415.

Turner, A. K., 1989. The role of three-dimensional geographic information systems in subsurface characterization for hydrogeological applications. In: Raper, J. (Ed.), *Three dimensional applications in geographical information systems*. CRC Press, pp. 115–127.

Watson, J. D., Stent, G. S., 1980. *The double helix: a personal account of the discovery of the structure of DNA*, 1st Edition. A Norton critical edition. Norton, New York.

Weisberg, M., Muldoon, R., 2009. Epistemic landscapes and the division of cognitive labor. *Philosophy of science* 76 (2), 225–252.
URL `http://www.jstor.org/stable/10.1086/644786`

Wible, J. R., 1994. Charles Sanders Peirce's economy of research. *Journal of economic methodology* 1 (1), 135–160.

Wilkins, J. S., 2008. The adaptive landscape of science. *Biology and philosophy* 23 (5), 659–671.

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the sixth international congress on genetics* 1 (6), 356–366.

Wright, S., 1994. *Molecular politics: developing American and British regulatory policy for genetic engineering, 1972-1982*. University of Chicago Press, Chicago.

Yuan, B., Gallagher, M., 2009. The Web Page of Max Set of Gaussians Landscape Generator. `http://itee.uq.edu.au/~marcusg/msg.htm`, Accessed 10 March 2013.

Zollman, K. J., 2010. The epistemic benefit of transient diversity. *Erkenntnis* 72 (1), 17–35.

Zuchowski, L. C., 2012. Disentangling complexity from randomness and chaos. *Entropy* 14 (2), 177–212.