Noname manuscript No.
(will be inserted by the editor)

Funding Science by Lottery

Shahar Avin

Abstract Motivated by recent criticisms of the low reliability and high costs of science funding allocation by grant peer review, the paper investigates the alternative of funding science by lottery, and more generally the possible introduction of a formal random element in the funding process. At first it may seem that randomness will lower expected efficiency, by allocating funds to less meritorious projects. By focusing on the notion that we want funded research projects to ultimately make our lives better, and the observation that the causal effect of research projects is subject to change over time, the paper argues that the introduction of randomness can counteract a bias towards the familiar present in grant peer review, and thus increase the overall efficiency of science funding. The time-dependant nature of scientific merit is exemplified by the historical processes leading to the discovery of the structure of DNA. The argument regarding the relative effectiveness of random allocation is supported by a computer simulation of different funding mechanisms on a hypothetical dynamic epistemic landscape.

Keywords science funding, grant peer review, random allocation, research funds, scientific merit

1 Introduction

Contemporary public support of basic scientific research is conducted primarily via allocation by peer review. Under this mechanism, researchers write descriptions of the projects they would like to pursue, and the proposals are ranked by their peers according to their perceived scientific merit. A ranking of the proposals is thus produced, and funding is awarded from the most meritorious downward, until the funds run out.

Recent empirical evaluations of grant peer review have raised concerns about its operation: Graves et al (2011) find it is not reliable, and Herbert et al (2013) find it is very costly. The findings have led Graves et al. to suggest a reconsideration

Shahar Avin
Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge, CB2 3RH, United Kingdom.
E-mail: sa478@cam.ac.uk

of an old proposal by Greenberg (1998), that will allocate a certain portion of the research funds to researchers at random. This paper presents a philosophical motivation for seriously considering the lottery option, by presenting a causal notion of scientific merit and arguing for difficulties involved in estimation of this quantity. The main source of difficulty considered is the dynamic nature of scientific merit, i.e. the possibility of significant changes in the merit of a research project over a short period of time.

2 Grant peer review

Grant peer review is the dominant contemporary mechanism for allocating public resources to basic scientific projects. Some aspects of the process are strongly conserved across nations and institutions (NIH, 2013; NSF, 2013; Dinges, 2005; Graves et al, 2011).

Grant peer review often offers significant 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.

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

Usually assessment is sought from more than one source, e.g. from multiple reviewers or from a mix of internal and external reviewers. 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. There are never enough resources to fund all projects. As such, comparisons of integrated assessments are used to decide which projects will get funded and which will not.

3 Empirical evidence for problems with grant peer review

Two recent empirical studies, presented below, look at the level of variability in the grant peer review decisions, and at the cost of running the peer review scheme.

3.1 Measuring the variability of peer review scores

How can the effectiveness of peer review be measured? One fairly good measure would be to compare the scores of reviewers to the impact of proposed projects (actual in case of funded projects, counterfactual in case of unfunded 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 (Dinges, 2005). The key limitations preventing this kind of study are the lack of information about the impact of projects which were not funded, and the absence of established indicators for measuring the impact of science.

A weaker evaluation of the validity of peer review scores is to check their consistency: to what extent different panel members agree among themselves about the merit of individual projects. The most thorough measurement published to date of the variability of grant peer review scores was conducted by Graves et al (2011). The authors used the raw peer review scores that individual panel members assigned to 2705 grant proposals submitted to the National Health and Medical Research Council of Australia (NHMRC) in 2009. In the original funding scheme, these scores were given within panels of seven, nine, or eleven reviewers, 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.

In their analysis, the authors resampled from the original scores to generate counterfactual scores. Thus, if the original scores were consistently low or consistently high, resampling will generate a counterfactual average score similar to the original average score. However, if the original scores featured a mix of high and low scores, the resampling will generate counterfactual average scores in a wide range of values. The authors used the counterfactual scores of each project to derive a score interval, or a range of possible scores that the project may have received had the panel composition been different.

The results of the study 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 authors claim the results show "a high degree of randomness", with "relatively poor reliability in scoring" (p. 3). The authors suggest further research, including investigating 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 difficult to discriminate. (Graves et al, 2011, p. 4)

The above quote suggests a clear link between variability in scores and a (limited) use of a lottery in funding. 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.

3.2 Measuring the cost of grant peer review

The cost of the grant peer review system can be broken down into three components: the cost of writing the applications, the cost of evaluating the proposals and deciding on which application to fund, and the administrative costs of the process. 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 exercise. 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 the NHMRC in March, 2012. The results of this survey, discussed below, are reported in Herbert et al (2013).

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.

The authors also conducted regression analysis on the survey results. 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 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 proposals were significantly less likely to be funded, when compared to new proposals.

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 merit of projects, such that high variability exists for a significant middle group. A possible response would be to accept an inherent uncertainty in the process, and cut costs by introducing a less reliable, but cheaper, allocation mechanism, such as a (limited) lottery, especially if some aspects of the current system already operate in a lottery-like manner. The next sections present a reasoned consideration of this alternative.

4 Worries regarding random allocation

There are some immediate objections that can be raised against the proposal to reduce the amount of merit evaluation in peer review and grant room for chance. The central worry is about effectiveness: if we do not rely on evaluation of merit, we would miss out on good research proposals, and will instead end up funding a lot of mediocre science. Challenging this worry will be the main focus of this paper. For completeness, another group of worries regarding the lottery proposal is discussed below, though these worries will not be treated at length.

An expected effect of greater randomness in funding allocation will be a change in the trajectory of research programmes. Under merit evaluation, it is often possible to receive continuous funding for a successful research laboratory, as long as new results are obtained and published and the technology and methods are considered cutting-edge. In contrast, under certain implementations of a lottery mechanism both successful, unsuccessful and novel programmes will have equal chances to win grants, and the relative portion of funds going to continuous funding will be reduced.

A cluster of worries can be associated with discontinuous funding. Continuous funding offers a measure of freedom that can entice highly-skilled individuals despite lower wages compared to other careers. A move to less continuous funding will lose this advantage and may result in less power to attract talent to science.

Many forms of scientific knowledge are gained slowly over time through practice, and are very hard to transfer efficiently to others. With discontinuous funding there is a real threat of losing this gained expertise and accumulated tacit knowledge, which will have a detrimental effect not just on individuals but also on the research environment.

Unexpected scientific discoveries can occur at any point during a funded research project. If a discovery is made close to the end of a funding period, under discontinuous funding there will be less scope to conduct follow-up research, reducing the payoff from such late discoveries.

Much of contemporary research requires significant infrastructure which is tailored to the research project. Such infrastructure needs to be set up, in a costly and time consuming process, any time a new avenue of research is initiated. Continuous research funding offers higher chances of reusing existing infrastructure, and thus offers an efficiency advantage over the costly set-up costs associated with discontinuous funding.

While these are all important worries, they are ultimately technical in nature, and may be solved using appropriate institutional design and practices that will complement the core proposal of funding by lottery. Not so the worry about efficiency, which is deeply associated with the core of the lottery proposal. The empirical evidence surveyed above suggests this might be less of a worry than originally envisioned, but further progress requires a more detailed conceptual analysis of effectiveness in science funding.

5 Scientific merit

The supposed advantage of grant peer review over random allocation is its ability to make approximately true comparisons between the scientific merit of alternative research projects. An evaluation of peer review's ability to make such comparisons reliably will require a working definition of scientific merit. But what is scientific merit?

A normative definition of scientific merit can be obtained from the initial rationale for public support of research. While the nature of the relationship between a society and its supported scientists may be complex (Geuna et al, 2003), the often cited motivation for public support of science strongly resembles the argument given by Bush (1945). According to Bush, public support of science leads to improvements in health, security, the economy and quality of life. To account for varying social preferences, a more robust definition is given:

Scientific merit (of a research project) is the extent to which the various causal consequences of the project contribute to well-being.

The above definition is deliberately left ambiguous with regards to the exact meaning, or measurement, of well-being. To state the point more formally, merit assignment $M(P, W)$ is a function that takes two parameters, the research project P, and a specific notion of well-being W, and assigns to them a merit score. The merit score of a project, given a certain notion of well-being, can be thought of as how close the consequences of the project will bring us to the specific notion of utopia that emerges from that particular concept of well-being. Thus, for a given notion of well-being, it is possible (in principle) to make comparisons between alternative research projects.

The definition of merit presented above is directly suggestive of some problems that will be involved in its measurement. It is tempting at this point to ditch the proposed definition and opt for a more tractable one. However, the definition as presented above captures a simple but significant notion, that we as a society devote non-negligible resources to scientific research *because we expect science to make our lives better*. Thus, rather than ditch the definition, let us be explicit about the worries of measurement, and follow them through to their consequences.

6 Difficulties in measuring scientific merit

A major difficulty in evaluating scientific merit, as will be argued below, is that merit evaluations are time-dependant. There are two closely related, but distinct, worries involved in this time dependancy. The first worry arises from partial and/or fallible knowledge about a target domain. Information about how a certain domain of research may best be explored is unlikely to be available in full until the domain has already been researched; thus, decisions about the most meritorious approach in a certain domain of research contain a substantial element of uncertainty, and future information gained from research may often show past merit evaluations to have been erroneous, despite relying on the best available information at the time.

A second worry is that domains of research are not static. Especially in domains where human and/or technological interventions are sought, such interventions may change significantly the character of the domain, while research is still ongoing. Thus, merit comparisons that rely on the domain being in one state may turn out to be false for the state of the domain at the time when research is being conducted or when the impacts of research are meant to take place.

In both of the above worries, the concern is that the choice to fund project A rather than project B, due to higher assigned merit, would, in hindsight, turn out to have been less effective than if project B was funded. In such a case we would have been misguided, or ignorant, in our assignment of merit, and the worry is that such ignorance may be pervasive. If such ignorance is pervasive, then the relative lack of effectiveness in lottery funding disappears, and with it the strongest objection to the proposal.

Two other difficulties with the evaluation of merit should be mentioned. First, the information about merit is diffuse. The full range of consequences of a research project play out in a wide arena, spatially, temporally, and contextually. Polanyi (1962) addresses this worry with regards to consequences within science, but a full evaluation of relative project merits will depend on knowledge of information diffusion, technological innovation, policy making, and other realms of expertise that may be far removed from scientific practice. It is a challenge facing the evaluators of projects to gather the necessary expertise required to meet this highly heterogeneous demand for information.

Another difficulty originates from the subjective nature of well-being. Like other public servants, science funding bodies are charged with making decisions on behalf of the public, and with a motivation towards the public's best interest. However, the aggregation of public preferences is a notoriously difficult task even in lay matters, let alone in preferences regarding scientific outputs that may require a significant level of tutoring before the preferences can be considered informed (Kitcher, 2011).

The above difficulties, regarding diffuse information and subjective evaluation, may turn out to be merely technical. Unlike the worry regarding time-dependant merit, these difficulties are not unique to science, and apply to other matters of public policy. More work is required to ascertain their significance for effective funding allocation, but such work lies outside of the scope of this paper.

Returning to the problem of merit changing over time, the next sections present an evaluation of the extent of the problem, and its consequences. The investigation proceeds in two stages: the next section presents a historical episode featuring multiple occurrences of rapid merit change, based on the account given by Allen (1975); the section following generalises from the historical example by means of a computer simulation.

7 Discovery of DNA: a historical example of rapid merit change

Two threads of the story of the discovery of the structure of deoxyribonucleic acid (DNA) 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.

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. In 1940 the Phage Group was started, 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.

In 1944, Oswald T. Avery provided the first direct demonstration that DNA was the genetic material. 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, the reception of Avery's 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. Given this changed biochemical background, a similar experiment to Avery's 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. 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.

The increasing interest in DNA, detailed above, led several groups to attempt to decipher its molecular structure. 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.

8 Modelling science funding under dynamic merit conditions

It might be argued that the historical episode described above is highly unusual in the history of science, involving a unique paradigm-shifting combination of events. To address this criticism, a model has been developed to capture the essence of dynamic merit changes, extrapolating from the example above to less dramatic, but more frequent, occurrences of merit change. The model is a variation of the epistemic landscape as constructed by Weisberg and Muldoon (2009). In Weisberg and Muldoon's model, a community of investigators sets out to explore a particular topic of interest. The various approaches to investigating the topic are represented in a two-dimensional configuration space, with distance between coordinates representing the similarity of the two approaches represented by these coordinates. Each coordinate (approach) is associated a scalar height (significance or merit), representing the value of pursuing that particular approach in investigation of the topic. The community of investigators performs well when the approaches of maximal merit are rapidly found and pursued.

To model dynamic merit, Weisberg and Muldoon's model has been modified to include time-dependant merit. This has been achieved by adding *trigger events*, such that when a particular trigger situation occurs, a change of merit (height) takes place in the epistemic landscape. In the simulation, three such trigger effects have been included:

– Following Strevens (2003), it is observed that little merit is associated with pursuing an approach that has already been successfully pursued in the past. Thus, whenever an approach is successfully pursued by an investigator, the merit of that approach is set to zero for the remainder of the simulation.
– Following Popper (1959), the value of a discovery is positively correlated with the amount of surprise it generates. Thus, when a significant discovery is made (an approach of merit beyond a certain threshold is first explored), nearby approaches lose some of their merit because they would now lead to less surprising results.

– Given the historical example above, it is clear that advances in one area can lead to new avenues of research in another area. In the simulation, this is represented by the appearance of additional merit in a random location on the landscape, following a sufficiently important discovery (when an approach with merit above a certain threshold is first explored).

In order to compare various funding mechanisms, the model has been further modified to include changes in the population of investigators over time. Three mechanisms which have been included are:

– *Best visible*: periodically, new entrants to the field propose to work on approaches at random locations on the landscape. A central funding body only considers those approaches which are sufficiently similar (near) to previously explored approaches, and selects from them the most meritorious (highest) candidates. Thus, as time progresses and familiarity with the topic increases, a wider set of approaches is considered viable, and merit selection has a wider pool to choose from. This mechanism was designed as a simple representation of grant peer review, where merit-based decisions rely on the past experience of experts.
– *Lottery*: periodically, new entrants to the field propose to work on approaches at random locations on the landscape. A central funding body chooses from them at random, regardless of the merit of their proposed approaches or whether they lie near or far from historical approaches.
– *Triage*: a combination of *best visible* and *lottery*, this mechanism supports half its candidates based on high merit from projects similar to historical approaches, and half its candidates by lottery from approaches which are dissimilar to past approaches. This mechanism was designed as a simple representation of the proposal by Graves et al (2011) mentioned in Sect. 3.1.

A visualisation of the simulation, including merit dynamics, is shown in Fig. 1 for the *best visible* funding mechanism, in Fig. 2 for the *lottery* mechanism, and in Fig. 3 for the *triage* mechanism.

9 Simulation results

The simulation was run on landscapes of various sizes, comparing the relative performance of the various funding mechanisms. The measure of success was the total accumulation of merit, i.e. the sum of merit from all pursued approaches throughout the duration of the simulation. The results are shown in Fig. 4 for a landscape of 50x50 approaches, and in Fig. 5 for a landscape of 200x200 approaches.

The results show that on the smaller landscape *triage* and *best visible* strategies outperform *lottery*, suggesting that for niche or restricted areas of research a peer review approach provides an advantage. In comparison, on the larger landscape the *lottery* and *triage* mechanisms outperform *best visible*, suggesting that in very open areas of research, or in situations where multiple topics can combine into one “super-topic” via interdisciplinary links, peer review loses its advantage and a lottery system becomes more appealing.

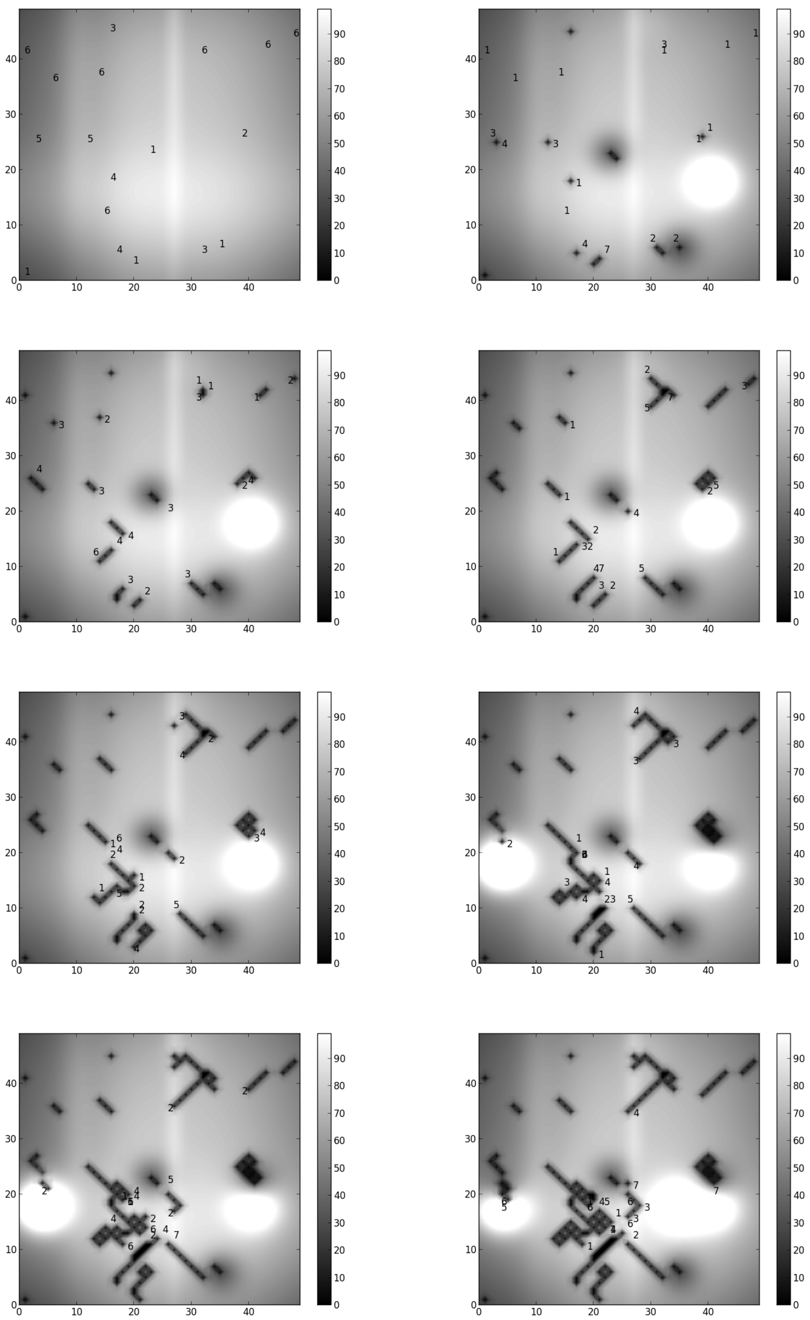

Fig. 1 Simulation of *best visible* funding mechanism on a dynamic landscape. Numbers represent locations of investigators, hue at a coordinate represents its height (brighter is higher).

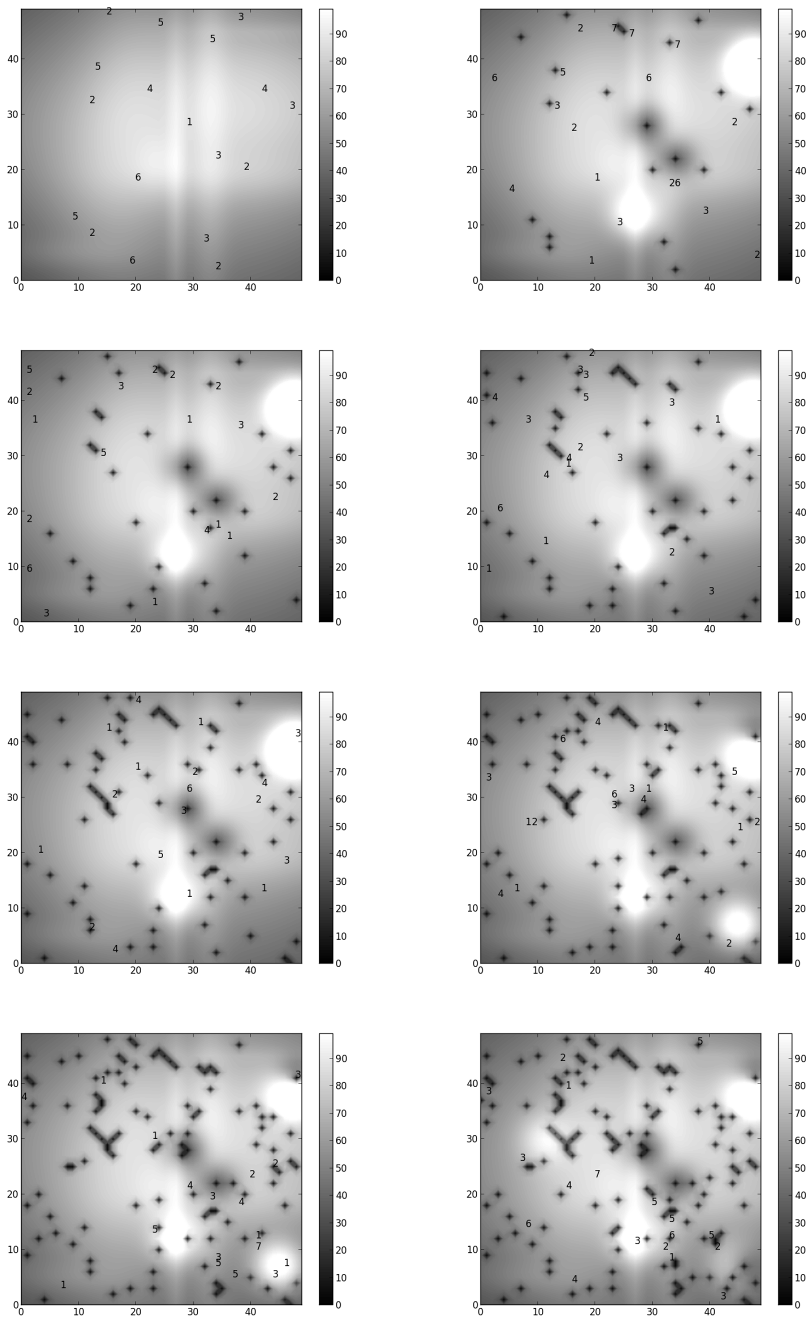

Fig. 2 Simulation of *lottery* funding mechanism on a dynamic landscape.

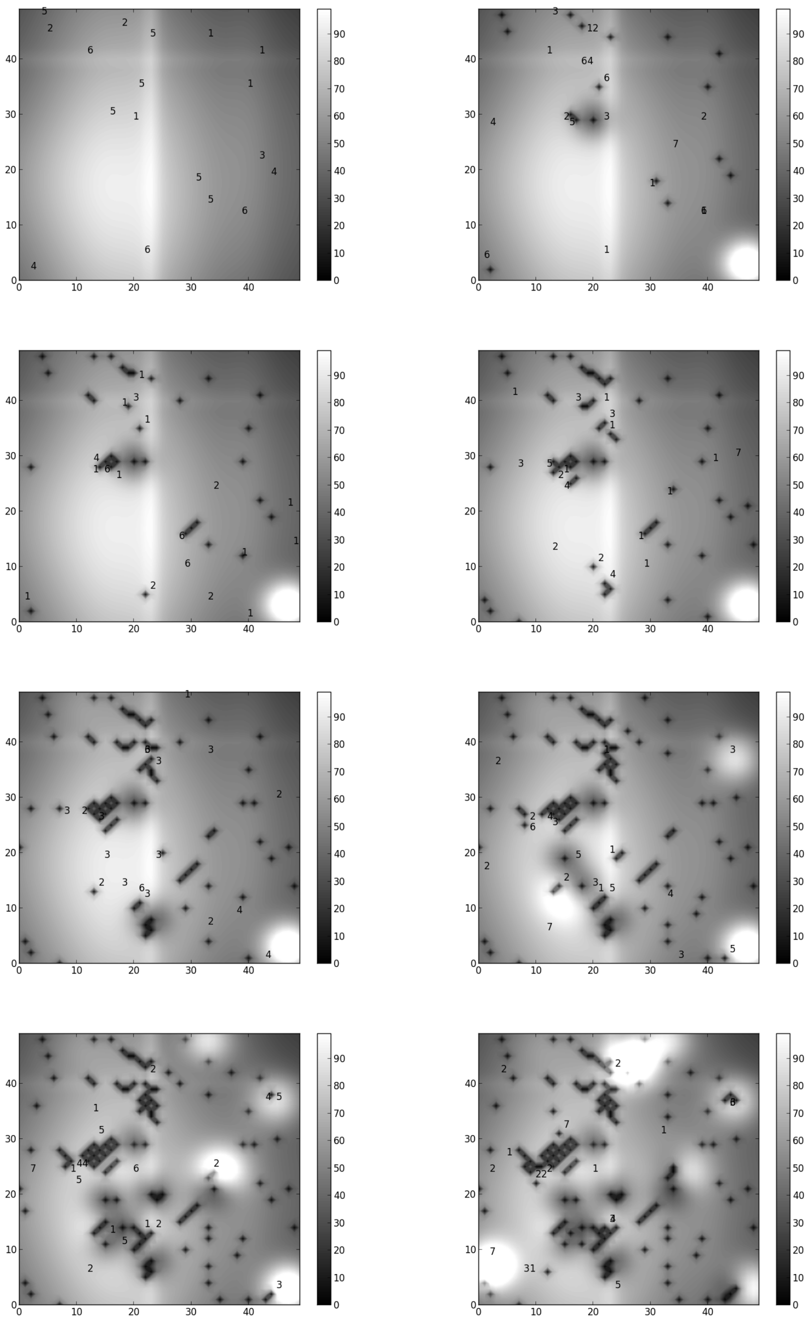

Fig. 3 Simulation of *triage* funding mechanism on a dynamic landscape.

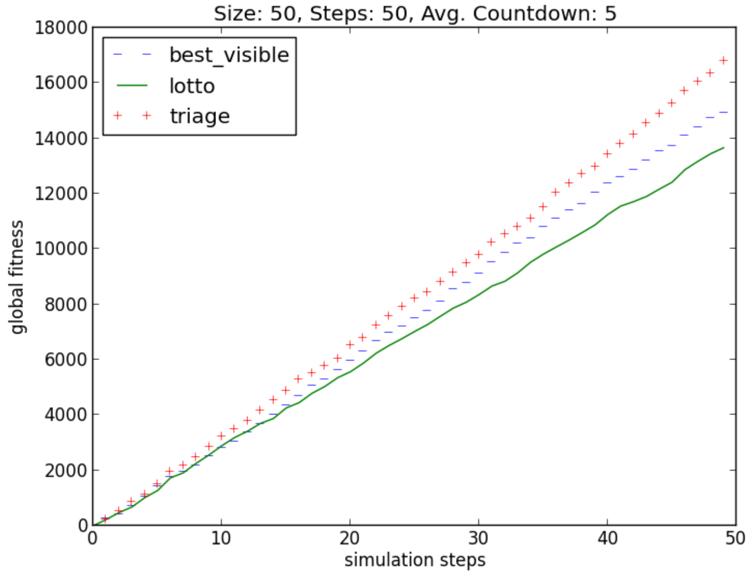

Fig. 4 Comparison of performance for different funding mechanism over time on a dynamic 50x50 landscape.

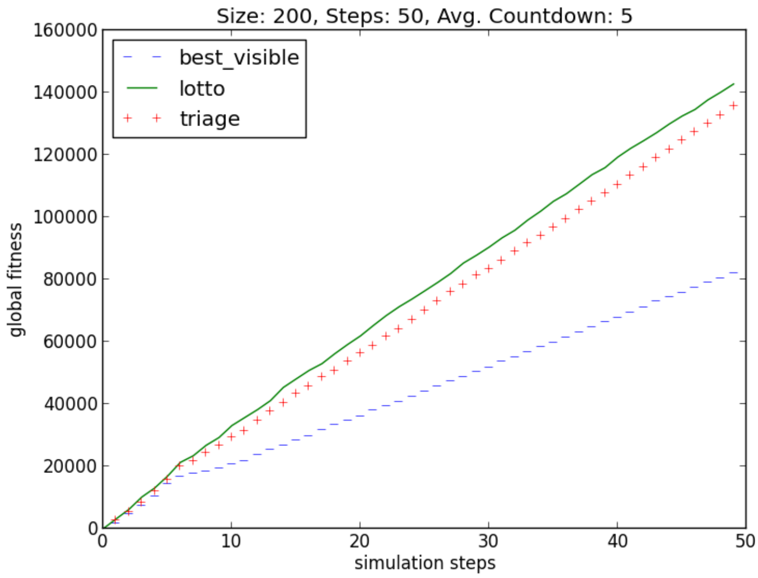

Fig. 5 Comparison of performance for different funding mechanism over time on a dynamic 200x200 landscape.

10 Discussion and conclusion

The simulation results given above flesh out a reasonable conjecture, that in wide and largely unexplored areas of research, past experience, and expertise that relies on past experience, is only of limited value. Given the drive within science towards exploration of the unknown and revision of the known, both empirically and theoretically, and the importance of connecting domains of knowledge via interdisciplinary research, it should not come as a surprise that grant peer review is becoming less reliable. The efficiency advantages of random allocation, which at first may seem absurd, are cast in a different light given this explication of reasonable assumptions we already hold regarding the advancement of science.

The relative advantage of the *triage* mechanism on both small and large landscapes suggests a happy medium, as this mechanism combines elements from both peer review and random selection. There could be various ways of implementing such a system in practice, but in all implementations two common features will be present:

- A formal randomisation element will be introduced to select from the pool of proposals amongst those whose merit evaluation is difficult or inconclusive.
- Less information and debate will be required for each proposal, because the exact merit scores of proposals which enter the lottery will no longer matter.

Such a system would reduce the overall cost of the funding exercise, while maintaining overall high effectiveness for scientific research. Rather than worry about lack of reliability in science funding, we should embrace it.

While sketching the core argument for formally including a random element in science funding above, many details of implementation and a discussion of possible consequences have been set aside for lack of space. A more thorough consideration of these matters is presented in Avin (2014), as well as source code for the simulations presented in the previous section.

References

Allen GE (1975) Life science in the twentieth century. History of science, Wiley, New York

Avin S (2014) Breaking the grant cycle: On the rational allocation of public resources to scientific research projects. PhD thesis (forthcoming), University of Cambridge, Cambridge, UK

Bush V (1945) Science, the endless frontier: A report to the President. U.S. Government printing office, Washington

Dinges M (2005) The Austrian Science Fund: Ex post evaluation and performance of FWF funded research projects. Institute of Technology and Regional Policy, Vienna

Geuna A, Salter AJ, Steinmueller WE (2003) Science and innovation: Rethinking the rationales for funding and governance. Edward Elgar Publishing, Northampton, MA

Graves N, Barnett AG, Clarke P (2011) Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel. BMJ 343, DOI 10.1136/bmj.d4797

Greenberg DS (1998) Chance and grants. The Lancet 351(9103):686, DOI 10.1016/S0140-6736(05)78485-3

Herbert DL, Barnett AG, Clarke P, Graves N (2013) On the time spent preparing grant proposals: an observational study of Australian researchers. BMJ Open 3(5)

Kitcher P (2011) Science in a democratic society. Prometheus Books, Amherst, N.Y.

NIH (2013) NIH grants policy statement. http://grants.nih.gov/grants/policy/nihgps_2013/, Accessed 9 November 2013

NSF (2013) Grant Proposal Guide. http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg, Accessed 9 November 2013

Polanyi M (1962) The republic of science: Its political and economic theory. Minerva 1:54–73

Popper K (1959) The logic of scientific discovery. Hutchinson

Strevens M (2003) The role of the priority rule in science. The journal of philosophy 100(2):55–79

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