Policy Considerations for Random Allocation of Research Funds

Shahar Avin

Penultimate draft. Final draft forthcoming in *Roars Transactions*.

Abstract

There are now several proposals for introducing random elements into the process of funding allocation for research, and some initial implementation of this policy by funding bodies. The proposals have been supported on efficiency grounds, with models, including social epistemology models, showing random allocation could increase the generation of significant truths in a community of scientists when compared to funding by peer review. The models in the literature are, however, fairly abstract (by necessity). This paper introduces some of the considerations that are required to build on the modelling work towards a fully-fledged policy proposal, including issues of cost and fairness.

Introduction

Proposal to fund science, at least in part, by random allocation, have been made both within philosophy of science (Gillies, 2014; Avin, 2015) and in other fields (Greenberg, 1998; Brezis, 2007; Graves et al., 2011; Fang and Casadevall, 2016). There are now at least three major funders who allocate a portion of their funds through a process that involves random selection: the Health Research Council of New Zealand's "Explorer Grants" (Health Research Council of New Zealand, 2017), New Zealand's Science for Technology Innovation "Seed Projects" (Science for Technological Innovation, 2017) and the Volkswagen Foundation's "Experiment!" grants (VolkswagenStiftung, 2017).

These policies are supported, at least in part, by modelling work (Brezis, 2007; Avin, 2017) that shows how introducing a random element to the funding process performs better than the current best practice of grant peer review, which allocates funds by relying entirely on expert judgement based on past experience.[1] These models highlight the value of random allocation in allowing greater exploration of the space of possible projects. At the same time, they portray tradeoffs between this increased exploration rate and the efficiency gains that come from relying on past experience and expert evaluation. Indeed, the main contribution of these models, it would seem, is in fixing the concepts required for evaluating these two causal processes, and the tradeoff between them. Unsurprisingly, they abstract away much of the remaining context of science funding policy. This paper, then, aims to fill in some of this missing context, so that policy makers and interested academics who are convinced by the key message of these models (or any of the other arguments supporting funding by random allocation) can start turning the proposals sketched by the models into actual policy recommendations.

In §1 I will review existing evidence for the cost and accuracy of grant peer review. In §2 I will review theoretical considerations relating to the use of lotteries in other domains (admissions and distribution of goods). §3 presents a proposal for how a science funding lottery might be run in practice, while taking into consideration desiderata and constraints raised by the previous two sections. In §4 I consider some limitations which define areas where funding by lottery is unlikely to be the best policy.

1 Empirical evidence for problems with allocation by peer review

The first step in bringing the random allocation proposal into the context of contemporary science policy is to ask what problems with current allocation mechanisms the policy may solve. The current dominant mechanism for allocating public funding to research projects is grant peer review, where proposals are invited from practising scientists, and these proposals are then evaluated by scientific peers for merit. Funding is allocated according to this

[1] These modelling results overlap, to some extent, with agent based models of publication peer review (Zollman, 2009; Thurner and Hanel, 2011; Squazzoni and Gandelli, 2013; Bianchi and Squazzoni, 2015), though the difference in context between grant peer review and publication peer review is significant, for example in the role played by uncertainty.

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 may improve the system by simultaneously increasing the eventual impact[3] is in contrast to treating impact only within the 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.

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. Nonetheless, Dinges is 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

[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).

[3]Given the context of public funding of science, I will use the term impact to mean the causal consequences of research that eventually (though possibly with much downstream effort and delay) contribute to social well-being, which I consider to be the core reason for public support of science. This is in contrast to causal effects that are entirely contained within academia, which are predominantly the ones captured by metrics such as number of citations.

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, 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. 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.[4]

- 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 of a

[4]Eugenio Petrovich has kindly pointed out to me that in response to this challenge, the field of bibliometrics has been developing normalised citation scores, for example the Mean Normalised Citation Score (MNCS) used by the CWTS Leiden Ranking (CWTS, 2017). However, such normalised indicators have also been criticised, e.g. by Leydesdorff and Opthof (2010). More generally, the need for dynamic indicators, and for caution in application across fields and for evaluation, are key tenants of the Leiden manifesto (Hicks et al., 2015).

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 scientific merit. 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, 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).[5] The authors used the raw peer review scores 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.[6] In this method, a set of bootstrap samples, often 1,000-10,000, are obtained from

[5] 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.

[6] For an introduction to bootstrap methods see Davison and Hinkley (1997).

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).

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). 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 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 existing modelling literature, one reasonable explanation would be that the variable scores are assigned to proposals outside of the past experience, or "vision range", of 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. An experiment run by Boudreau et al. (2016), in which "vision distance" was directly measured, suggests that the effect of increased conceptual distance is to introduce *bias* rather than *uncertainty*, with degree of uncertainty remaining roughly constant, and similar in magnitude to that found by Graves *et al.* Boudreau et al. (2016) broke down "vision distance" into two components: "evaluator distance", the degree of content similarity between a reviewer's area of expertise and the area of the proposal, and "proposal novelty", the degree of content similarity between the proposal and all known works in the area of the proposal. They used overlap of standardised keywords, assigned by an independent librarian, to measure these distances.

The above quote from Graves *et al* 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.[7]

[7]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

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.

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 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.

---

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).

Using survey data, the authors also tried to detect a correlation between extra time spent on a proposal and the proposal's likelihood of being funded. Surprisingly, no such correlation was found, and given the power of the study this suggests that, on average, 10 extra days spent on a proposal are likely to *at most* increase the likelihood of success by 2.8% (p. 3). The authors did find a statistically significant correlation between the probability of success and whether the proposal was a resubmission of a previous (failed) proposal: resubmissions were less likely to be funded, on average, when compared to new proposals.[8]

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. 1, in black).

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.

[8]The authors do not provide a hypothesis to account for this observation. We could hypothesise that a significant portion of failed proposals represent low merit projects within the visibility range of the scientific discipline. Since, over a short period of time, significant gain of scientific potential is more rare than significant loss of scientific potential (as the field progresses it "exhausts" the area of familiar projects), what is once labelled as low merit (if within the vision range) is likely to be similarly labelled in subsequent years, until a rare breakthrough re-infuses the exhausted field with new potential.

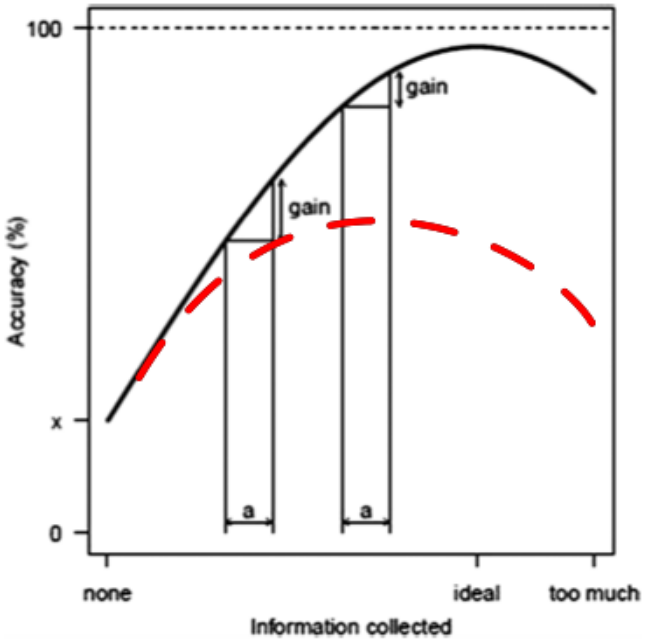

Figure 1: The accuracy of peer review assessment as a function of information provided. Original figure, in black, is reproduced from Herbert et al. (2013, Fig. 2, p. 5), and represents the authors' hypothesis, not a conclusion from their data. The red dashed curve was added by me, and represents an alternative dependance. Herbert et al. (2013) was published under CC-BY-NC licence: `http://creativecommons.org/licenses/by-nc/3.0/legalcode`.

The authors rely on their result, that no statistically significant correlation was found between extra time spent on a proposal and its likelihood of success, to argue that the current amount of information provided is more than 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 about reviewer's accuracy (Gillies, 2014), 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. 1, dashed red curve).

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.[9] The prevalence of lotteries and their unique features have generated various theoretical works in political theory, economics, and moral philosophy.[10]

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.

[9] 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).

[10] Books of note on the topic of lotteries include Boyle (2010); Duxbury (1999); Gataker and Boyle (2008); Goodwin (2005); Stone (2011).

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. Boyle's definitions are:

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 who 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.

Boyle proposes the following example of how a lottery might have been introduced into a selection mechanism to make it more fair. 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. The eleven plus IQ test was considered the most reliable predictor of the five-year academic success of students out of the available measures, though it was known to be imperfect. 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%.[11] 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, 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

[11] 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.

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.

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 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, it is both more efficient and more fair to pick the projects of highest merit, because merit takes into account the information needs of the entire population. Nonetheless, when comparing mechanisms of equal ability to generate scientific value, 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 and models in the literature suggest it will be large. 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, as the information demands and values of the society change. Nonetheless, we can use the measurements of variability discussed in §1 as a guideline for setting up the "border zone" for grant proposals.

2.1.1 Criticisms and responses to Boyle's paper

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.

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.[12] Thus, if we compare two candidates, the error in the test would

[12]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,

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 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.

e.g. in the order of the questions. As discussed earlier in the chapter, there is no good mechanisms for empirically uncovering general systemic bias in peer review results. Where specific biases are detected, e.g anti-novelty bias (Boudreau et al., 2016), measures can be taken to address them, but that still leaves the possibility of further undetected biases.

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 to the academic success of children who went to less-academic schools). 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.

- 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.

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 models of science funding, and bears very strong resemblance to Gillies' argument against the homogeneity-inducing effects of peer review.

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.[13] In a weighted lottery, no matter how bad your chances are, you have at least some chance of winning.

[13]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).

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.

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. 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 paper 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.

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.

2.2.1 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.

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, i.e. whether individuals how are highly motivated to do research end up producing better research.

- 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.

2.2.2 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 highly uncertain and dynamic nature of science significantly undermines 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, some element of auction (in the form of promise competition), and motivation for exaggerated promise, already exists in the current peer review system.[14]

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. Unless other solutions could be found, an auction-like mechanism seems to be ruled out for science funding.[15]

[14]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).

[15]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.

2.2.3 Distribution by queue or evaluation of 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. 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.

The issue of the expected utilisation of research funds rules out a first-comes-first-served queue model for science funding. Given the evidence presented in §1, that extra time spent on a proposal does not correlate with higher likelihood of success, it is unlikely that Boyce's system of earned merit is a good model of science funding applications, though at a coarse grained level we may say that the high time investment involved in grant applications leads to self-selection amongst scientists. There is, however, no clear reason to believe that the scientists most able and motivated to spend significant time on grant applications are those most likely to maximise utilisation of grants.

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 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.

2.2.4 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 replaced with an activity that achieves something of value to the community, for example contribution of time and experience to the education system or relevant industries, or mentoring young researchers. If mixed with some light checking mechanisms (e.g. those proposed by Gillies (2014); Fang and Casadevall (2016)), we get a system that guarantees some minimal level of utilisation, reduces lost costs (by reducing the time wasted on detailed applications), introduces rebate (in the form of "participation fees") and, according to Boyce's analysis, increases participation.

3 Design of a possible science lottery policy

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". 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 in the design of a science funding policy.

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 US National Science Foundation (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), rather than, say, academic disciplines or addressed social need, as communities engaged in a particular epistemic activity are best positioned to accumulate and access knowledge regarding the relevant skill set and similar past projects. 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 (Polanyi, 1962), 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.

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.

There is insufficient space here to defend the second point, but 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.[16]

[16]This point about situation-dependent personal traits bears strong resemblance to the situationist account of moral character presented by Doris (2002).

- The relevance of some of these skills depends on the specific nature of the research project, but 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, or that it is better to craft policies as if there are no such cases, these special provisions may be dropped.

3.3 Use short proposals to locate projects in the space of possible projects

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 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 impact, and some extrapolation into the future of the field and the expected impact of the proposed project. Since information provided in the proposal is slim, the assignment should be rough, into groups of "known high merit", "known medium merit", "known low merit" and "unknown merit". It

might be possible to introduce graduation within the unknown merit 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, where the application should be funded outright.[17] Preferably, the main bulk of the detection of exceptional talent 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. Either way, this would only capture a small subset of as-yet-undetected talent, as one reviewer noticed. In cases where individuals have already demonstrated exceptional talent through a major contribution there should be, and are, available funding streams outside the lottery to support them.

3.4 Triage proposals, using a lottery for the middle group

The assignment of expected value, based on the location of the project in the space of possible projects, is used to triage the proposals:

1. All proposals of known high merit should be funded. Based on the results of Graves et al. (§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 merit and proposals of unknown merit should be placed in a lottery. If graduation is used for the unknown merit group, a graduated lottery may be used accordingly, in a similar manner to Boyle's graduated lottery.

[17] 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 false positive rate for such cases may be quite high, 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.

3. All proposals of known low merit 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 merit 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 merit, 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 merit group) a chance of funding, and will also help prevent over-specialisation of the domain. The high merit 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 merit peak.

3.5 Managing potential outcomes of introducing a lottery

There may be initial upheaval following the introduction of random selection into a hitherto 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 that uncertainty in research is ineliminable, and a limited lottery has a good chance of yielding better results for society in the long run.[18]

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

[18] 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.

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 merit, and therefore if that merit is high they would be funded without a lottery, and if that merit is not high then perhaps the loss is not so great. Furthermore, best practices could be devised for documentation, facility swap, and skill transfer, so that the costs of freezing and unfreezing projects is lowered.

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.

4.1 Very expensive projects

The lottery mechanism was designed with a certain (common) project size in mind: projects that 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 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.[19]

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 paper, 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).

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.

[19] Current funding practices also sometimes fail in providing sustained support for mega-projects, for example the partially-funded Superconducting Supercollider in the USA.

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 paper, that of the public support of basic research, such cases are not the norm.

4.3 Fully explored area

When an area of research is known to be fully explored, the space of possible projects 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.

4.4 Researcher identity determines scientific impact

The value 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 value 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 value of the project, because the researcher's authority influences the effect the research will 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 value 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 impact 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

Theoretical models and arguments have focused on efficiency when suggesting that random selection may outperform other mechanisms for choosing research projects. This paper goes beyond the theoretical models and looks at other desiderata for a funding mechanism, mainly fairness and cost, showing how they can be taken into account in the design of a lottery-based funding mechanism. To do this, the paper surveyed existing evidence about the empirical reliability and costs of science funding exercises, considered issues of fairness in reference to a case of a lottery in an education setting, and considered the lottery from the perspective of the applicants using an analysis from the economics literature. These were all combined to create a more detailed and nuanced template for science funding by lottery than what exists in the theoretical literature. Further analysis was presented regarding areas where a science funding lottery is unlikely to provide benefit, and may instead cause harm.

Early implementations of random allocation are poised to provide empirical evidence about the payoffs of funding by lottery and about the reactions of different stakeholders and publics to the policy. Until such data become available, and given the heterogeneity of environments in which random allocation might be implemented, we should continue exploring different ways to organise scientific activity.

References

Avin, S., 2015. Funding science by lottery. In: Recent Developments in the Philosophy of Science: EPSA13 Helsinki. Springer, pp. 111–126.

Avin, S., 2017. Centralised funding and epistemic exploration. *British journal for the philosophy of science*.

Bianchi, F., Squazzoni, F., 2015. Is three better than one? simulating the effect of reviewer selection and behavior on the quality and efficiency of peer review. In: Winter Simulation Conference (WSC), 2015. IEEE, pp. 4081–4089.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., Riedl, C., 2016. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science* 62 (10), 2765–2783.

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.

Brezis, E. S., 2007. Focal randomisation: An optimal mechanism for the evaluation of R&D projects. *Science and Public Policy* 34 (10), 691–698.

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.

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.

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.

CWTS, 2017. Leiden ranking. http://www.leidenranking.com/; last checked 4 November 2017.

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.

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.

Fang, F., Casadevall, A., 2016. Research funding: the case for a modified lottery. mbio 7: e00422-16.

Frazier, S., 1987. *University funding: Information on the role of peer review at NSF and NIH*. US General Accounting Office.

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.

Gillies, D., 2008. *How should research be organised?* College Publications, London.

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., 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.

Health Research Council of New Zealand, 2017. Explorer grants. `http://www.hrc.govt.nz/funding-opportunities/researcher-initiated-proposals/explorer-grants`, last checked 22 June 2017.

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).

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., Rafols, I., 2015. The leiden manifesto for research metrics. *Nature* 520 (7548), 429.

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`

Kitcher, P., 1993. *The advancement of science*. Oxford University Press, New York.

Knuth, D. E., 1997. *The art of computer programming*, 3rd Edition. Addison-Wesley, Reading, MA.

Latour, B., 1987. *Science in action*. Open University Press, Milton Keynes.

Leydesdorff, L., Opthof, T., 2010. Normalization, cwts indicators, and the leiden rankings: Differences in citation behavior at the level of fields. *arXiv preprint arXiv:1003.3977*.

Martino, J. P., 1992. *Science funding: politics and porkbarrel*. Transaction Publishers, New Brunswick, NJ.

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

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`.

Saunders, B., 2008. The equality of lotteries. *Philosophy* 83 (03), 359–372.

Science for Technological Innovation, 2017. 2017 SEED PROJECT FUNDING Q & As. `http://www.sftichallenge.govt.nz/sites/default/files/2017-06/2017%20SEED%20funding%20Q%20and%20As_0.pdf`, last checked 7 September 2017.

Squazzoni, F., Gandelli, C., 2013. Opening the black-box of peer review: an agent-based model of scientist behaviour. *Journal of Artificial Societies and Social Simulation* 16 (2), 3.

Stone, P., 2011. *The luck of the draw: The role of lotteries in decision-making*. Oxford University Press, Oxford.

Thurner, S., Hanel, R., 2011. Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B* 84 (4), 707–711.

VolkswagenStiftung, 2017. Experiment! – in search of bold research ideas. `https://www.volkswagenstiftung.de/en/funding/our-funding-portfolio-at-a-glance/experiment.html`, last checked 22 June 2017.

Zollman, K. J., 2009. Optimal publishing strategies. *Episteme* 6 (2), 185–199.