Draft: Feb 5, 2022

Christian Hemmestad Bjerke, chrbjerke@gmail.com
Marco Ottaviani, marco.ottaviani@unbocconi.it
Kristin Oxley, krox@forskningsradet.no

Funding Science in an Uncertain World: The Role of Information and Incentives[1]

Uncertainty is the second nature of science. One cannot know whether years of research will ever lead to a discovery, whether the scientific community will realize the significance of the discovery, and whether the discovery will lead to important applications. COVID-19 vaccines were eventually developed at breakneck speed for a global pandemic that, as of January 2022, has caused more than 5.5 million deaths (World Health Organization 2022), and the development of a new vaccine class which holds great potential for the future (Pardi et al. 2020). The development of mRNA vaccines is a story of remarkable scientific progress in an environment shrouded in deep uncertainty (see also Franzoni et al. 2021).

The thrill of discovery, the traction of recognition and prestige, as well as the pursuit of career incentives such as job security and tenure were all important drivers that led scientists to develop the fundamental research behind the vaccine. Pulled by the urgent need for action, various streams of research came together at an unprecedented speed. The story behind these scientific endeavors underscores the delicate interplay between the motivation of scientists and the availability of funding for research institutions, large pharmaceutical companies, and startups.

The mRNA genetic molecule, a key component of COVID-19 vaccines, was discovered more than 60 years ago (Cobb 2015). A traditional vaccine works by injecting a modified version or selected components of the virus into a patient's body, thus training the immune system to learn how to combat the virus and apply the lesson in the future if needed. The mRNA genetic molecule works differently by carrying instructions on how to produce the viral proteins directly to the patient's cells (CDC 2022). This creates a more realistic infection that trains the immune system more effectively, with future benefits in case of contraction of the disease. However, for years the use of mRNA molecules for vaccine development was thought to be largely impractical. Given their very fast degradation, it seemed almost impossible to isolate mRNA molecules from cells. The focus of molecular biologists shifted to other, more promising areas of research (Kolata and Mueller 2022).

Scientists like Dr. Weissman and Dr. Karikó believed in the potential of the mRNA molecule as a carrier of instructions. Their beliefs, however, did not fall in place with the then prevailing paradigm, and they faced limited funding (Rothberg 2021). Unfazed by this and convinced of their hunches, they spent seven years researching how to inject mRNA without provoking an adverse immune response. Once they managed to solve the problem, the

[1] We are grateful to Federico Pessina for truly outstanding research assistance.

scientific community did not realize the importance of the discovery. There were still many obstacles, such as the molecule's fast degradation when injected into a host which meant that mRNA was still considered an unrealistic candidate for vaccines.

Key breakthroughs in devising an efficient method for preventing the molecule's degradation were developed in parallel by other scientists working in both academic and commercial settings. The interplay between these environments is a common feature of research, and it can lead to conflicts. Another team of scientists found a way of encasing the mRNA molecules in a lipid shell. While promising, there were a number of technical difficulties, and the technology was licensed to Protiva, a company that made great strides in improving efficiency and perfecting the approach (Jeffs et al. 2005). Dr. Karikó understood the importance of the work and approached Protiva for collaboration. Discussions around intellectual property and the demands of the funding parties prevented this collaboration from occurring (Kolata and Mueller 2022).

Breakthroughs by Dr. McLellan, Dr. Graham, and Dr. Wang in visualizing coronaviruses and developing a method to stabilize Spike proteins were another essential component of the Covid-19 vaccine. Through a sequence of serendipitous events, Dr. McLellan's and Dr. Graham's lab realized that a common coronavirus could hold lessons for its more deadly cousins which cause MERS and SARS. The first breakthrough came soon when the team managed to isolate and visualize the virus's Spike proteins, an unprecedented feat (Kirchdoerfer et al. 2016). The lab, in particular through Dr. Wang's work, achieved a second breakthrough by devising a method to stabilize the Spike proteins, a serious issue that had prevented the development of vaccines (Pallesen et al. 2017). Again, the breakthrough was overlooked in the scientific community. The MERS epidemic was over and coronaviruses were not an area of focus (Kolata and Mueller 2022).

The importance of these breakthroughs was understood only in the first months of 2020 with the emergence of COVID-19. Suddenly, the research of hundreds of scientists that had been disregarded became of utmost value. Dr. Weissman and Dr. Karikó have been recognized with a Breakthrough Prize, the largest award in science, and are considered strong candidates for the Nobel Prize (Dolgin 2021).

An essential role in the development of mRNA technology was played by research-intensive startups like Moderna and BioNTech, founded in 2008 and 2010 respectively with the support of both public and private grants. At a later stage, large established pharmaceutical companies such as Pfizer entered the field, enabling the roll out of the vaccine.

The critical situation that emerged in the first months of 2020 led governments and institutions such as the World Health Organization to adapt and revise their standard approach in managing the deployment of vaccines. In a race to roll out the vaccine, governments entered advance purchase agreements with vaccine producers well before the vaccines were approved by the relevant national health regulatory bodies (Bloom et al. 2021). In July 2020, the U.S. government purchased 100 million doses of the Pfizer-Biontech vaccine for $1.85 billion, with the option of acquiring 500 million more doses in the future (Pfizer 2020). Similar contracts were stipulated by Israel, the United Kingdom, and the EU. Advance purchase agreements are an unconventional form of pull incentives. The aim is to

stimulate and expedite firms in their final stages of innovation by providing them with significant market rewards upfront.

This swift response was made possible by well-positioned startups that took decisive and risky actions at the very beginning of the pandemic, and then formed strong ties and partnerships with established pharmaceutical companies. Within days of the virus' genome sequence being available online, Moderna had created a prototype vaccine. Within 8 months from early human testing, Pfizer and Biontech received emergency authorization.
This previously unheard of pace of development, testing and approval, was made possible by hands-on management and partnerships between companies, research labs, and governments.

Uncertainty Reigns

Science is a journey into the unknown to push the frontiers of knowledge. Research continuously challenges and refines what we believe is true. Scientists always entertain the possibility that the existing body of knowledge is incorrect or at least incomplete. Gleick (1992) writes that Feynman "believed in the primacy of uncertainty, not as a blemish on our ability to know, but as the essence of knowing." When facing unsolved problems, "you have to leave the door to the unknown ajar" (Feynman 1998).

The lack of certainty and the pursuit of improved knowledge is what spurs scientific innovation. Certainty is impossible to achieve as our understanding of the world evolves. Rather than certainty, we should reasonably expect reliability from science. Scientific ideas, in both physical and social sciences, are reliable and credible when they survive the continuous arrival of new evidence and the repeated scrutiny of the scientific community. But they are not certain (Rovelli 2012).

According to Kuhn (1962), science evolves with alternating stages of normal science and scientific revolutions. Normal science operates within an accepted paradigm, which consists of a set of tools, methods, notions, and assumptions that are largely shared by the community of scientists working in a field. Scientific revolutions take place when a critical mass of evidence unveils some fundamental inconsistencies in the prevailing paradigm. As novel tools, assumptions, and theoretical frameworks are developed, a new paradigm emerges that is more compatible with the evidence. Revolutionary science permits science to move between different stages of understanding. This model of scientific progress evokes the key role of uncertainty as described by Feynman (1998) and Rovelli (2012).

We can immediately see how a scientist whose work does not fit within the current paradigm and, of course, is not certain of being part of the next paradigm, faces significant uncertainty in their professional lives. Practical aspects of a scientist's life, such as job security and achieving tenure, for example, are heavily influenced by the scientist's scientific productivity and standing in the scientific community. The distribution of citations within scientific fields is heavily skewed. It is clear that scientists face significant uncertainty when deciding which research avenue to pursue. Each choice has a high opportunity cost and scientists cannot easily predict the best path to take. Wang et al. (2017) construct a measure of novelty for individual publications and find more novel publications tend to have higher variance in their citations and take longer to be recognized as valuable by the scientific community.

Another aspect of uncertainty arises from the potentially delayed responses to scientific research. It is not always the case that advancements in science are immediately recognized as useful. The discovery of the mRNA molecules is an excellent example. The fast degradation of mRNA molecules led to the premature conclusion that it was not a viable area of research. Molecular biologists moved their focus to other rising areas of research, such as DNA and proteins (Kolata and Mueller 2022). A minority of scientists believed in the molecule's potential and, despite the uncertainty and risks involved, managed to solve the issues that had stunted progress. It took several decades for the significance of the discovery of the mRNA molecule to be vindicated. How can "sleeping beauties", papers whose importance is not recognized for several years after publication (Ke et al. 2015), be identified early?

In a recent study, Menkveld et al. (2021) uncover another important source of uncertainty that is generated by the latitude researchers have in analyzing the data, even when both the data to be used and the hypotheses to be tested are given. They distinguish the data-generating process (when samples of data are obtained from a population) from the evidence-generating process (when estimates from the sample are obtained through statistical analysis). While research has typically focused on the standard errors that arise from drawing a sample from a population, the evidence-generating process can also cause "non-standard" errors, which add uncertainty to the scientific process. Non-standard errors may arise from the choices that researchers must make when analyzing the data, from the data cleaning method to the specification of a statistical model—even when the data and the hypotheses are held constant. The set of choices is vast and it can lead to sizable errors. In an experiment where 164 teams tested six hypotheses on the same data sample, Menkveld et al. (2021) found that non-standard errors have a similar magnitude to standard errors, are routinely underestimated by the researchers involved, and tend to be reduced by peer feedback.

Science as Incentive System

Driven by passion and intellectual curiosity, investigators carry out "blue-sky" research to solve open puzzles and challenge conventional wisdom. Often it is not clear whether the potential discoveries will lead to practical applications. Scientists operate within a community whose shared values and ideals have been studied extensively by sociologists of science. According to Merton (1942), the spirit of modern science is captured by four sets of institutional imperatives: communalism, universalism, disinterestedness, and organized skepticism. These norms are not upheld necessarily for their intrinsic ethical value, but because they are considered by the majority of the profession as essential for the renewed success of scientific research. Scientific ideas have universal value, are common property and are subject to the scrutiny epitomized by the motto of the Royal Society, *Nullius in Verba.*

The science reputation system rewards the first scientists who discover important new knowledge and present it with clarity to the scientific community. Rather than assigning the property of the innovation to the discoverer, the priority system rewards the publication of results. The discoverer essentially donates the innovation to the community, in exchange for public recognition. Once the innovation becomes common property, other scientists can

freely build on it, while properly crediting the contribution by citing the relevant work. Citations act like non-monetary dues, thus making the priority reward proportional to the value of the contribution for follow-on work and discoveries.

Scientists are largely driven by a set of complementary incentives such as prizes, awards, and the tenure system. These nonmarket incentives leverage the scientists' intrinsic motivation to make discoveries and to gain prestige and reputation for scientific excellence. These complementary incentives reinforce each other as a system in which each element can only be understood in combination with the other elements (Holmstrom and Milgrom 1994).

Science, like any other human endeavor, is far from a perfect system. Incentives sometimes are insufficient, other times excessive, and often distorted. To be promoted and be granted tenure researchers are under pressure to obtain funding, deliver measurable results, and publish work that attracts citations. In addition to picking low hanging fruits, researchers might have an incentive to avoid upsetting the common wisdom of reviewers and journal editors, or worse, to cut corners and adopt questionable research practices, in stark contrast with the Mertonian norms. These practices challenge the very credibility of science. In a number of fields, many empirical studies have proven to be difficult or impossible to replicate, even when using the same dataset, leading to a replicability crisis. A key driver of this crisis is publication bias, resulting when only significant results or results that are aligned with partisan views of either the researchers or the reviewers are published.

While the motivation of scientists to thrive in the scientific community plays a central role in the arena of pure research, commercial incentives are also essential for innovations to come to fruition and be widely adopted—think of the deployment of the new vaccines to the world population. Two systems of incentives are relevant for practical innovation, the nonmarket "science system" and the "market system". The internal organization of science is far from the market system traditionally analyzed by economists. The science system is largely governed by nonmarket incentives, mostly acting through non-anonymous interactions that build on scientists' self-motivation for discovery and interest for recognition, impact, and prestige. Market transactions, instead, are less reliant on individual personalities and center more on impersonal and anonymous interactions typically mediated by monetary rewards and the price system.

Market incentives play a central role in the diffusion and deployment of applications of scientific knowledge. As new knowledge is embodied in innovation, the commercial adoption of practical applications is largely driven by market incentives. Incentives centered on intellectual property such as patents provide monetary rewards to innovations that deliver consumer value. Policy levers are typically classified into push and pull incentives:

- Push incentives, such as grants, credit, and tax subsidies that lower research costs, act on the cost/supply side at the ex ante level. Push incentives are typically unconditional and leave investigators free to pursue their curiosity. Researchers are thus empowered to follow hunches that might lead to break-throughs. However, push incentives run the risk of being technology driven and untargeted toward applicable innovations that are widely adopted.

- Pull incentives, such as intellectual property, instead act ex post through the revenue/demand side, and thus tend to be more goal oriented. The promise of market rewards offered by patents, for example, incentivize firms to bring their products to commercial fruition. These rewards, being conditional on outcomes, are highly targeted, but are often perceived to be insufficient to motivate fundamental research that can have far-reaching applications that are difficult to predict and trace back to the source.

While push incentives tend to play an important role in the early stages of innovation where basic research is prevalent, pull incentives typically become prevalent in later stages of commercial development. To be effective, science policy must understand the interplay of science and the market, and the delicate nexus of incentives that prevail in the two systems. Prizes, advance market commitments, and goal-oriented research are more elaborate forms of incentives that contain elements of conditionality typical of traditional pull incentives, but they act ex ante, similar to push incentives. These mixed tools hold great promise but require careful governance.

1. Science Policy as Organizational Design

How does or should society organize science? Economists naturally frame the production and diffusion of knowledge as an organizational design problem that builds on social choice and agency theory, a paradigm that has become central in modern economic analysis since Arrow's (1951 and 1963) pioneering work. Principal-agent problems arise from the simultaneous presence of two features:

(1) Misaligned Interests: Individuals have conflicting preferences (objectives, motivations, and interests). For example, scientists (agents) might enjoy doing research on intellectually challenging problems, more than on problems that are practically relevant for the funder (principal).

(2) Asymmetric Information: Individuals make decisions under uncertainty, having access to limited information, which is typically distributed asymmetrically across individuals. For example, scientists might privately know better than the public the potential benefits and costs of a line of research.

For the sake of argument, at least initially, let us abstract away from the complications created by the fact that principal and agent each represent many individuals. For the grand scheme of science policy we can think of the agent as the collective of scientists and the principal as the research funding organization (representing society).

Focus on the baseline formulation with a single agent and a single principal. The principal aims at controlling the behavior of the agent through a number of contractual and organizational tools. Assuming feature (1) and in the absence of informational frictions (2), the divergence of objectives of principal and agent can be framed in terms of externalities. Externalities measure the impact of the action taken by the agent on the payoff of the rest of society. Under symmetric information, traditional public economics suggests how the externality problem can be solved through a corrective tax and subsidy policy (Pigou 1920). The idea is simply to compensate the agent with a payment equal to the externality.

Activities that generate negative externalities, such as pollution, should be taxed, while activities like research that generate positive externalities should be subsidized.

At the other extreme, suppose that principal and agent have access to asymmetric information as in (2), but they have aligned preferences, so that feature (1) is absent. This case corresponds to the economic theory of teams, pioneered by Marschak and Radner (1972). Team theory primarily deals with the aggregation of information across individuals and the communication problem that is also studied in computer science.

Agency problems arise from the simultaneous presence and interplay of conflicts of interests (1) and asymmetric information (2). The two main specifications of asymmetric information are (a) hidden action (i.e., moral hazard) problems in which the decision made or effort chosen by the agent is not observed by the principal and (b) hidden information (i.e., adverse selection) problems in which the principal does not observe some key parameter (or "type") that characterizes the agent. For example, in a moral hazard problem the funding organization might not observe how hard the researcher is really working to solve the relevant problem, while in an adverse selection problem the funder might not know the intrinsic merit of the researcher's idea.

In the case of science, researchers has large latitude in the choice of action because of the high level of uncertainty in the environment. Direct monitoring is rather ineffective, given the specialized knowledge required. And there is a fundamental informational problem in knowing which problems should be investigated. The science incentive system leverages the self-motivation of researchers, their quest for knowledge, the thrill of discovery. Incentive problems would be limited if researchers could self finance their activities, like the independently wealthy gentlemen scientists impersonated by Robert Boyle (Shapin 1995).

Modern science, however, requires large resources—it is more an industrial activity than an independent art. Proper funding is needed to attract talented individuals from all walks of life. At the same time, funding risks distorting scientists' incentives as some would be attracted by the monetary rewards rather than the more intrinsic pleasures of scientific research. Successful funding schemes should leverage the self motivation of scientists, while directing efforts toward socially valuable goals. Information asymmetry constrains this process. Instruments such as peer review in the allocation of grants are attempts at reducing this asymmetry.

Modern research universities are major funders of research. According to the prevailing model, scientists perform both teaching and research. It is often claimed that teaching revenue is largely used to cross-subsidize research (James 1990). Promotion and tenure criteria heavily weigh research performance. A possible explanation is that teaching quality is hard to measure when student evaluations are sometimes akin to popularity contests which hardly reflect added value. Research performance, instead, is more accurately observable through publications and citations. When teaching and research abilities are correlated, it becomes efficient for principals to use research performance as a selection device.

2. Funding Instruments

Scientific research is an arduous and demanding activity that requires significant resources. Scientific progress has been at the root of economic growth in the modern era. Modern

governments acknowledge science policy issues as having strategic importance for national competitiveness and economic security (Dasgupta and David 1994). In this context, an investigation of the economics of the funding of science is of critical importance.

The main product of science is invention and discovery of new knowledge. Arrow (1962) explains why we should not expect the market system to deliver efficient production and diffusion of new knowledge in the form of information. Information has a number of features that naturally lead to market failure. Firstly, information is non-rival: it can be replicated at zero (or close to zero) cost. Secondly, the production of information is risky. As discussed in the previous section, uncertainty is a key component of the scientific process. Not all research ventures lead to the production of knowledge. In addition, there are uninsurable moral hazard issues. Lastly, trading of information is problematic as the value of information is only revealed after it is transferred.

Because of the public good nature and positive externalities associated with research, we expect the market to underprovide innovation. For this reason, the role of the government in incentivising innovation and providing funding for scientific research leading to innovation is justified on economic terms. Governments, universities, public bodies, institutions and individuals have funded research through a number of instruments.

Prizes and Awards

Prizes consist typically in monetary rewards given to the scientist or group of scientists that manage to achieve a specific, predetermined objective or multiple objectives as defined by the funding committee. The Longitude Prize is probably the most famous example. In 1714, by passing the Longitude Act the British Parliament promised a payment equivalent of £1.5m in today's currency to whoever could devise a method to determine longitude to an accuracy of half a degree. The prize was created as a response to the high risks that ships and their crews faced in traveling across the Atlantic (Sobel 2007).

Prizes tend to be created in response to a well-defined need. Therefore, scientists have a clear goal to aim at. Prizes can only be effective when the final use of the innovation is clear at the outset. A main difficulty may arise in identifying the need and defining the goal in a precise way. Prizes are more open-ended than grants, as noted by Kalil (2006). Grants typically require the funding agency to specify the recipient of the award and the preferred approach. Prizes typically do not have these limitations and thus can more effectively incentivise innovation through unconventional means.

However, researchers who aim to win the prize are not awarded funds ex ante, as it would be the case with other funding instruments such as grants. Given that prizes do not make funds available until after the discovery is made, prizes can only be effective if researchers can leverage funding through other sources. Thus, entry is restricted only to candidates who have significant resources ex ante to support their research. Given the uncertainty involved, it might be difficult for inventors to finance their research activity. Obtaining credit from capital markets is difficult given the highly specialized expertise needed to evaluate the promise of research ideas. Indeed, funding from traditional credit markets is rarely available given the uncertainty and information asymmetry that characterizes scientific investigations. Difficulties with funding might explain why the use of prizes is rather limited. Mission-oriented

research schemes and advance market commitments can be seen as a modern incarnation of prizes in that they both direct research efforts toward clearly defined goals.

Awards differ from prizes as they have an important honorific component and are assigned ex post for outstanding achievements. The criteria for awards is typically specified in a more general manner compared to prizes. Awards satisfy the motivation for prestige by publicly acknowledging, celebrating, and enhancing the reputation of a scientist.

Patents

A patent confers to the inventor the property right to exclude others from selling the protected invention. By limiting ex post competition this way, a patent holder is free to exercise market power and charge prices that more than cover variable costs. The lure of profits attracts potential innovators who can expect to cover the ex ante fixed cost associated with the inventive activity. However, market power results in a deadweight loss of social surplus. Also, while the patent system in principle is designed to require the patent holder to disclose information about the invention, to maintain their market power patent holders might resist and obstruct followup innovations by competitors.

The patent system requires an ex ante evaluation of the claims of innovators (centralized at the level of the patent office) as well as an ex post mechanism to manage disputes (mostly decentralized through patent litigation). Inventors are largely guided by their own interest in the pursuit of new market needs. Inventors are successful only if they bring to market products or processes that buyers appreciate. Given that inventors must contribute their own resources and seek financing in the marketplace, the patent system is highly decentralized.

A significant difference between patents and grants is the award size. Grants entail a fixed reward. In the case of patents, rewards are essentially set by the market, depending on the relative demand and supply for the product covered by the patent (Hemel and Ouellette 2013). From an informational perspective, patents exploit the ex ante informational advantage the inventor has regarding the future value of the invention (Wright 1983). Patents are a flexible instrument because the market reward is linked to the user value of the innovation. By making the returns to the innovation private, innovators can leverage future profits to obtain private funding and finance their activities in the credit market.

Research and Development (R&D) Subsidies

R&D subsidies provide financial assistance to firms in an effort to stimulate innovation. The common approach is a tax subsidy. Governments provide tax advantages to R&D activities, thus implicitly reducing the cost of innovation.

Similar to patents, R&D tax subsidies are decentralized. No central planning authority specifies which ideas are worth investing in. However, general mission-oriented plans might assign subsidies to different sectors. Firms are free to allocate their time and resources to the most promising avenues of research. As these subsidies do not typically cover all the research costs incurred, incentives to recoup the investments through the market and to select the best ideas are not distorted.

Advance Market Commitments (AMCs)

AMCs are a novel form of funding instrument consisting of a funder committing to fully or partially finance purchases of a product which meets predetermined conditions. AMCs were utilized for the first time in 2007 for the development of the vaccine for pneumococcal disease (Kremer et al. 2020). In this case the funder agreed to purchase a quantity of vaccines in the future at a specific price. Vaccine developers are therefore guaranteed a return on their investment, conditional on the vaccine conforming to the conditions previously stipulated.

Low purchasing power of less developed countries has hindered vaccine development over the past decades. Pharmaceutical companies have historically shied away from investing in vaccine development due to the low profits expected. The aim of AMCs is to specify upfront a price and design a market so that companies have guaranteed demand for their product in the short run and can ensure a sustainable supply in the long run (Albright et al. 2005). By appropriately subsidizing the market price, AMCs allow vaccine developers to recoup their investment costs. AMCs exploit the informational advantages of vaccine developers, leaving them free to pursue the development of the vaccine in the most efficient way. Subsidies tend to be proportional to market sales, thus leveraging the information of buyers and encouraging effective and large-scale product deployment. This commitment is particularly important for vaccines considered technologically distant, for which funders may have particularly scarce information regarding the best approaches (Kremer et al. 2020).

Grants

Grants can be seen as temporary employment contracts, accompanied by a budget for expenses. Typically grantmakers give wide latitude to the grantees to spend the budget in the most appropriate way.

Grantmaking has its roots in the patronage system. Since time immemorial, governments and wealthy patrons have sponsored talented artists, renowned philosophers, and brilliant mathematicians and scientists. The great mathematician Archimedes was likely to be on the payroll of the government of Syracuse, where his inventions, especially in the military arena, were greatly appreciated. Philip II, King of Macedon, retained Aristotle to educate his young son, the future Alexander the Great. Galileo Galilei obtained an appointment as professor without teaching obligations—the dream of any academic—at the newly created University of Pisa by leveraging his connection with the young Grand Duke of Tuscany Cosimo de' Medici, whom he tutored as a child (Biagioli 1990 and Westfall 1985).

The process of grantmaking has been institutionalized over the last two centuries. In the nineteenth century, the French Academy of Sciences gradually developed its celebrated system of prizes into a system reminiscent of "encouragements" or grants, where relatively small monetary budgets were awarded to promising scientists to enable them to make new discoveries rather than to reward them retrospectively for having already achieved important innovations (Crosland and Gálvez 1989).

Hanson (1998) traces the dominance of grants over prizes to institutional features of the patrons that were supporting science societies and academies in the eighteen century. He

finds that societies with non-autocratic, non-local government patrons were especially likely to use funding instruments akin to grants rather than prizes. Grants eventually prevailed over prizes, as science funders became increasingly democratic and centralized.

Following the successful experience in government-funded research during WW2, the previous director of the wartime Office of Scientific Research and Development Vannevar Bush lobbied effectively for stronger federal research support. In 1950 the National Scientific Foundation (NSF) began its operations. In 2021, the NSF had an annual budget of $8.5 billion and supported around 25 percent of all federally supported basic research in US colleges and universities (NSF 2022).

Prizes involve a reward at the completion of an objective. Grants instead provide the funding ex ante, thus ensuring that researchers have the resources necessary for their studies. There is no guarantee that the project will achieve the promised objectives. For this reason, the initial stage of evaluating the grant proposals is of critical importance.

In a criticism of the distortions associated with using grant peer review for allocating science funding, Roy (1985) forcefully proposes an alternative system based on a formula that rewards research departments—rather than individual scientists—for the quality and quantity of their research output (such as publications and citations), while also taking into account some key research inputs. The key advantage of this alternative system, according to the proposer, would be its reliance on retrospective performance measures, which are more reliable than the evaluation of speculative research proposals. Research funding schemes like the UK Research Excellence Framework essentially follow these ideas.[2]

3. Grantmaking

Having introduced the main funding instruments, this section zooms in on the economics of grantmaking. Our discussion of the economic logic of grant funding complements Azoulay and Li's (2021) recent overview of the fast-growing empirical work in the area.

Retrospective Evaluation

Maurer and Scotchmer (2004) analyze the role of retrospective evaluation in grantmaking by formulating a simple infinite horizon model. Their model stresses how grant funding self-selects more productive researchers who value repeated funding more and thus have stronger incentives to deliver the results promised in the grant proposals they submit.

When awarded a grant, a researcher obtains an immediate private benefit equal to υ, corresponding, say, to career advances. If the researcher works hard on the project and spends effort cost κ, society obtains a benefit θ. Researchers differ in terms of their productivity, captured by the parameter λ that represents the probability they will obtain a fundworthy research idea in each future period. Researchers are privately informed about their productivity. The funder, instead, only knows the distribution of the productivity in the population of researchers. Researchers receive a grant for each idea.

[2] See Geuna and Martin (2003) for an overview of the issues and Reale and Zinilli (2017) for a more recent discussion.

In each period, the researcher must decide whether to work hard on the project, thus producing the expected research results, or to shirk saving the effort cost κ. In equilibrium, the grantmaker starts off in the first period by assigning an initial grant to every new applicant. In subsequent periods, however, the grantmaker will continue funding only the researchers who made good on their past promises and delivered results for all the grants awarded in the past. The researcher has an incentive not to shirk in the first period if

$$v \leq v - \kappa + \frac{\lambda}{r}(v - \kappa).$$

By shirking (on the left-hand side), the researcher saves on the cost κ, but loses the net benefit from all future grants, equal to per-idea rent $v - \kappa$ obtained by the researcher, times the probability λ of obtaining an idea in any given period, in perpetuity from next period (and, thus, divided by the discount rate r). This no-shirking condition can be rearranged as

$$\lambda \geq \frac{r\kappa}{v - \kappa}.$$

Only highly productive researchers have the right incentives to produce results and thus obtain continuous funding.

This model hinges on the simultaneous presence of hidden action and hidden information in grantmaking. First, the researchers have private information about their productivity type (hidden information). Second, there is asymmetric information about research effort (hidden action). The grant contract leaves researchers rather free, given how difficult it is to verify the specific activities carried out by researchers and the specialist knowledge required. Researchers are tempted to "take the money and run". Incentives to deliver in Maurer and Scotchmer's (2004) model results from the loss of the rent associated with future grants, which act like an "efficiency wage". More productive researchers value this rent more because they are more likely to obtain grants. As the model highlights, grantmaking is both an incentive device (giving researchers incentives to perform in order to obtain future grants) and a selection device (sorting more productive researchers into funding, while less productive researchers who have less stake in future funding yield to the temptation of not delivering).

Prospective Evaluation

Maurer and Scotchmer's (2004) view stresses the retrospective evaluation of the performance of grant applicants. While past performance surely plays an important role in grant funding, evaluation of applications also has a prospective role. Grant programs thoroughly evaluate and review proposals made by applicants. Only the applications deemed to be most promising are funded—for an indication of how competitive the process is, the current success rate (fraction of funded applications) at the European Research Council (ERC) is around 10%. Given that preparing a grant application is a time consuming and costly activity,[3] it is natural to expect that candidates should be more willing to apply when they expect to be more likely to be funded.

[3] In an observational study, Herbert et al. (2013) report that submitters of new proposals to the Australian National Health and Medical Research Council spent on average 38 working days completing the application.

To analyze the impact of prospective grant evaluation on the extent of the self-selection of candidates, Ottaviani (2020) proposes a model of grantmaking with noisy evaluation. The baseline version of the model with a single field can be understood in terms of demand and supply for funding. On the demand side, candidates are characterized by a continuously distributed type θ, which captures their intrinsic merit. A candidate bears a cost equal to c to apply; the applicant then obtains benefit v if awarded a grant. On the supply side, the grantmaker observes a noisy signal for each application submitted.

On the supply side, a higher signal indicates that the application has higher merit. The grantmaker then assigns the available budget of grants to the applications that obtain a signal or, equivalently, a score above a certain acceptance cutoff. In turn, on the demand side, candidates with higher merit are more likely to be awarded a grant. Thus, in equilibrium only candidates with merit above a certain application cutoff find it worthwhile to apply. How does the size of the budget available for distribution and the accuracy of the grant evaluation process affect the extent of self-selection of candidates?

An increase in budget reduces self-selection and increases the incentives to apply. But while applications increase with the budget available for distribution, whether the success rate increases or decreases with the budget depends on the shape of the distribution of researchers' types. Suppose that researchers' types have a thicker tail than exponential, consistent with a classic observation by Lotka (1926) that researchers' productivity in terms of publications follows a power law. Intuitively, applicants who self-select into applying are then much stronger on average than the marginal applicant, who is just indifferent whether to apply or not. As the budget increases, the additional applicants that are attracted are much weaker than the average inframarginal applicants, thus pushing the success rate down. In this case, an increase in the budget results in a more than proportional increase in applications, so that the elasticity of applications with respect to the budget is greater than one. This prediction is confirmed by evidence of the impact of the 2009 increase in the budget available for research grants in the US as a result of Obama's Stimulus Package to buffer the great recession. Grant applications increased more than the budget, thus resulting in a reduction in the fraction of successful applicants; see Stephan (2012, p. 145).

Ottaviani (2020) also shows that an increase in noise of the evaluation signal decreases the extent of self-selection and unambiguously increases the incentives to apply. Intuitively, as the evaluation becomes noisier, the probability of succeeding in obtaining a grant becomes less responsive to the merit type of applicants. Thus, more candidates must apply to make sure the applicants absorb the given budget of grants that is available.

While the allocation of budget within a panel representing a single field is relatively straightforward when expert reviewers are able to compare the merit of applicants who work in their specialized area, the apportionment of budget across different fields is much more delicate. For example, it is natural to expect that reviewers understand better and tend to favor work in their own area of expertise over work in other fields—a version of homophily bias.

If the determination of the final allocation across fields is left to a composite panel of experts from different fields, disagreement and conflict is likely to result. It is natural to expect experts to inflate the scores they assign to work in their field and to correspondingly

downgrade work in other, competing fields. The quality of the evaluation will suffer, resulting in a less reliable judgment.

Alternatively, the overall budget could be apportioned across fields using an automatic formula, as it is practically done at some of the world's largest research funding organizations. For example, the National Institutes of Health (NIH) and the ERC apportion the budget in proportion to funds requested by applicants in each panel. Proportional budget allocation automatically equalizes the success rate across fields. The system appears fair and balanced and indeed performs well if fields are relatively symmetric in terms of the noise in the evaluation signal.

If, instead, fields have varying levels of evaluation noise, the performance of proportional allocation deteriorates. Recall that, for any given level of budget, noisier fields tend to attract more applications. But with proportional apportionment the increased number of applications results in a proportional increase in the budget, which in turn induces a further increase in applications. If applications increase less than proportionally with the budget, as is the case when the type distribution has a thinner top tail than the exponential distribution, the process converges to an interior equilibrium that definitely features more applications in the noisier field. When the distribution of types is exponential in all fields, or has a top tail thicker than exponential, even the slightest asymmetry in noise results in all the funds being allocated to the noisiest field, with no applications in all other fields—a drastic version of unraveling. The performance of proportional apportionment can be truly perverse. Unraveling results more generally when evaluation in a field is sufficiently accurate; it always occurs in the extreme case with perfect evaluation.

4. Peer Review

We are now ready to zoom in on another prototypical agency problem in science—the delegation of evaluation to experts.[4] Society (the principal) should aim at selecting the most promising among a number of projects proposed by different scientists (the agents). Scientists benefit from funding. In the first stage, the agent proposes their project to the funder. There is informational asymmetry between the principal and the agent regarding the quality of the project. Applicants have strong incentives to selectively reveal information regarding their project in order to maximize their chances to be funded.

Peer review is a system used to minimize this informational asymmetry. Rather than directly evaluating the proposal, the funder delegates evaluation to a new set of agents, the peer reviewers. This way, the principal has two sets of agents, the scientists and the reviewers, structured hierarchically. Peer reviewers are selected for their expertise in the specific scientific fields relevant to the proposal. Their role is to reduce the information gap between the principal and the scientists.

The introduction of the reviewer leads to new agency problems. First, performing an accurate review requires not only expertise in the field, but also time and hard work. What

[4] See David (1998, 2008) for an engaging account of the historical emergence of the review process and the open science movement. For an extensive review of the scientometric literature on peer review we refer to Bornmann (2011). See also Rietzschel et al. (2019) for an overview of the organizational psychology literature on the assessment of creative work.

are the incentives of reviewers to perform an informed and balanced assessment? Reviewers are selected on the basis of their ability and willingness to act as evaluators. Naturally, reviewers who have research interest closely aligned to the work to be evaluated are more likely to be interested in reading the application and might be willing to contribute constructive comments.

At least in principle, the grantmaker (or more concretely, the program manager or other panel members) can try to estimate the expertise of the reviewers on the basis of their historical track record. When the grantmaker must simultaneously estimate the merit of the projects as well as the expertise of reviewers evaluating the very same projects, the problem becomes challenging. A complicating factor is that the grantmaker does not have access to exogenous information about the merit of the projects—and observes at best a noisy signal about the quality of projects, typically only when the projects are actually funded.

In practice, the review process leading to the selection of applications proposed for funding unfolds through a number of stages. According to a common organizational design, in a first stage proposals are reviewed individually. Then reviews are shared among panel members, who discuss and compare proposals at a meeting and reach a consensual ranking. Often the final allocation is determined either by a higher-level committee or a formula (e.g. proportional). There are many variations on this basic format; for example, first-stage reviewers might be panel members or outside reviewers; in some cases, outside reviewers are added between the first and the second stage.

Here we review the main biases and pitfalls that might affect the grant peer review process and how organizational design can be leveraged to address these issues.

Reviewer Bias

According to Merton (1973) the norms and values to which peers hold each other are universally and consistently applied to all members. When these norms and values are impartially interpreted and applied, peer evaluations are understood as being fair. Bias can be defined as the violation of impartiality in peer evaluation (Lee 2013).

Should we expect reviewers to report truthfully his/her opinion? Or should we be concerned about reviewers being biased and tilting their evaluation? A reviewer with specialized expertise closer to the applicant's field might be able to better evaluate the proposal, but might be biased in favor or against the application.[5] If reviewers stand to gain from work in their area (or that cites their own research) they might be overly supportive. If, instead, reviewers perceive a competitive threat they might be excessively critical and try to boycott the proposal.

In their review of the literature on peer review, Guthrie et al. (2018) find that there is some evidence (though not consistent across studies) that peer review can be biased in terms of gender, age, ethnicity and 'elite' vs peripheral institutions, and that it suffers from cronyism and cognitive particularism (favoring your own field or way of thinking). Regarding cognitive particularism, Travis and Collins (1991) find that reviewers tend to favor proposals supporting their own school of thought, and argue that this is likely to have a much bigger

[5] Garfagnini, Ottaviani, and Sørensen (2014) model this situation as a problem of mechanism design without transfers and with limited commitment.

impact on the direction of science than institutional bias or cronyism identified by other studies. Using NIH data, Li (2017) documents the same pattern; reviewers with research interest closer to the applicant are more favorably biased. However, they are also better informed. Other studies find that reviewers are more critical of applications in areas of their own expertise (Boudreau et al. 2016 and Gallo et al. 2016).

Cognitive particularism might be partially explained by the extensively documented information processing bias termed *confirmation bias*, which is the tendency to search for, interpret and recall information in a way that supports what we already believe. Information that confirms our expectations is more easily processed and understood, and thus has a bigger impact than does disconfirming information. Testifying to the subconscious nature of this bias, Lamont (2009) concludes based on her ethnographic study of peer review panels that “equating ‘what looks most like you’ with ‘excellence’ is so reflexive as to go unnoticed by some.”

Related to cognitive particularism is the finding that reviewers suggested by the applicants themselves tend to give higher scores. Analyzing data from the Australian Research Council, Marsh et al. (2008) conclude that applicant-nominated reviewers gave “biased, inflated, unreliable and invalid” marks, systematically higher than experts selected in other ways. Severin et al. (2020) confirm this finding in a recent cross-sectional analysis of 38,250 external peer review reports submitted to the Swiss National Science Foundation.

Langfeldt (2002) finds that professional assessments are not only colored by the interests of the assessors, but also by their different assumptions and thought patterns. Langfeldt and Moen (2013) point out that "while professional bias can give structural biases in the outcome of application processing—in the form of prioritizing certain disciplines, methods or fields—personal bias often has a more random effect.”

Heterogeneous Leniency

There is extensive empirical evidence that reviewers disagree to a high degree in their assessments of grant proposals. Jayasinghe et al. (2001, 2003) and Marsh and Bazeley (1999) suggest that an important source of grant reviewer disagreement is idiosyncratic response biases in the way assessors map their subjective impressions onto a numerical rating scale.

Lee et al. (2013) identify a number of studies which show significant differences in the patterns of reviewing by gender, with female reviewers being stricter than their male colleagues. They also find that toughness may vary by disciplinary affiliation and nationality.

Lamont (2009) find that such varying degrees of harshness and leniency might be partly explained by strategic voting whereby scoring “is guided primarily by a desire to facilitate or hinder the funding of another proposal” and can involve so called low-balling, “giving a lower rank that would otherwise be justified to some proposals” and high-balling, “strategically assigning a high rank to a proposal.”

Author Anonymity

Should the identity of applicants/authors be disclosed to the reviewers? Reviewers might be inclined to evaluate proposals submitted by prominent scholars or by submitters who belong to dominant groups more favorably, giving them an undue advantage. As discussed, there is for example some evidence that grant peer review is biased in favor of 'elite' vs peripheral institutions. The impact of double blind submission policies on discrimination has been studied extensively; for classic studies in economics, see Blank (1991) and Goldin and Rouse (2020).

When the author's identity is concealed, ex post sorting by reviewers might be hampered because the decision cannot be conditional on observable characteristics and other predictors of success. However, as Taylor and Yildirim (2011) highlight in the context of a simple model, anonymity levels the playing field among those submitted, thus increasing the incentives of strong and weak authors to perform. From a practical point of view, however, implementing double blind submissions and retaining author anonymity is increasingly difficult in the age of the internet.

Reviewer Anonymity

According to a widely adopted practice, the identity of peer reviewers is not revealed to submitters. A common explanation is that reviewers would otherwise be less objective because they would give in to social pressure to be supportive and avoid expressing candid criticism. For example, reviewers could be concerned that if they are openly critical of the work submitted, the submitter would hold a grudge against them and that their own work in the future would receive similar critical comments.

Anonymity might improve objectivity but, at the same time, anonymity may be counterproductive. For example, emboldened by anonymity, reviewers might take advantage of their power position. It might also reduce accountability. Name-Correa and Yildirim (2019) develop a theoretical model where reviewers have expertise in assessing a particular dimension of the proposal and are negatively biased toward this dimension. Reviewers might be worried about being singled out as the cause of an eventual rejection of the proposal. When the concern for blame is high, the principal can benefit from imposing anonymity so that reviewers don't excessively lower their acceptance standard. If, instead, the concern for blame is low, public voting becomes preferable.

Reputational Concerns

Suppose that reviewers are not biased, but they want to appear to be well informed. Would reviewers concerned about their reputation for expertise have the right incentives to truthfully report their evaluation?

Consider a game of reputational cheap talk with the following structure, based on Ottaviani and Sørensen (2006). Each of a number of reviewers privately observe a signal about the merit of a grant proposal. Then, each sends a recommendation to their peers in the panel. The panel uses the recommendations of all reviewers not only to assess the merit of the grant proposal but also to assess the expertise of the panel members. The objective of each reviewer is to maximize this assessment that the panel makes about their individual expertise.

At first blush, one might think that a reviewer interested in developing a reputation for accuracy should have the right incentives to report the posterior expectation about the merit of the proposal, conditional on the noisy signal received. However, a careful analysis of this game of reputational cheap talk reveals the somewhat counterintuitive conclusion that experts who aim to maximize their reputation do not necessarily have an incentive to be truthful.

When more informed reviewers have signals closer to the merit of the applicant, reviewers would like their peers in the panel to believe that their private signal is located at their posterior expectation about the candidate's merit, in order to convince the panel that their private information is accurate. In other words, if reviewers can convince peers that their predictions are based fully on private information, they would be considered even better informed than they really are. However, in equilibrium, the panel should adjust for this incentive of each reviewer to shade their recommendation toward the consensus view. The upshot of this reputational incentive is that reviewers cannot credibly reveal all the information they have. See for example Nunnari et al. (2018) for an experimental investigation of reputational cheap talk in the context of strategic forecasting.

In some cases this loss of information can be drastic. For example, when the picture emerging from other reviewers is rather clear, individual reviewers who might have information leading in the opposite direction might keep silent. This tendency to conform reduces the accuracy of the evaluation process.

Aggregation of Reviews

Panels aggregate the information conveyed by different reviewers and reach a decision regarding the proposal. The process of aggregating peer reviews has informational and incentive implications, and thus must be designed with care.

Reputational concerns may lead to herding. A typical approach is a sequential process where each agent reveals their information in turn. Imagine a situation in which each agent has received a signal regarding the quality of a proposal and are called to share their information in a public debate. As the process unfolds, agents are learning from the statements of the other agents, and thus will continuously update their beliefs about the quality of the proposal. Given that agents are worried about appearing well-informed, they may decide to not reveal their true signal if it's in disagreement with the statement of a reviewer with better expertise. Potentially valuable information is discarded. This leads to herding, where agents selectively reveal their information and decide to 'follow the crowd' (Ottaviani and Sorensen 2001).[6]

The order in which reviewers are asked to give their opinion is also important. This is due in part to the anchoring bias (Kahneman 2011), which refers to the strong significance of arbitrarily set anchoring points. In line with this theory, introductory comments often serve as a reference point for all subsequent comments. Research on peer review indicates that this

[6] Banerjee (1992) and Bikhchandani et al. (1992) develop the logic of herd behavior and informational cascades in a related setting in which decision makers act—rather than speak—in sequence. When a sufficiently large number of predecessors make the same decision in a common value environment, it becomes optimal for individuals to completely disregard their own information and take the same action as the predecessors.

also holds true in grant peer review discussions. For example, according to Arensberg (2014), “what the first speaker starts the discussion with is found to be decisive . . . The strong or weak points of the applicant mentioned first are strongly supported by other panelists.”

A common sequencing approach is the anti-seniority rule where experts speak in order of seniority, which typically is a good proxy for expertise. This rule is used for example in many judicial systems in which judges are promoted to high level courts only once they demonstrate sufficient expertise. The benefit of this approach is that more junior experts, such as junior judges, might be less influenced by their more senior colleagues. While the anti-seniority rule is not always optimal, it tends to be more attractive than the seniority rule according to which more informed experts are consulted first.

Some funders address the issue of herding by requesting that all panel members supply written input prior to the panel meeting in the form of scores and/or comments and structuring the discussion around the information supplied.

The importance of aggregating reviews cannot be understated. Herding can have disastrous consequences. The tendency of groups to suppress dissent has also been studied extensively by social and organizational psychologists. Recounting a number of policy disasters, Janis (1982) argues that groupthink emerges when cohesive decision-making groups establish a mode of thinking that prioritizes consensus rather than a more rational appraisal of the alternatives.

Even though the literature on expert aggregation contains some useful insights, much work is left to be done to obtain concrete predictions on how the deliberation protocol can be designed to maximize the quality of the evaluation process.

Shared Information Bias and Conformity Bias

The tendency to conform to the majority view and the tendency for groups to focus on information known to all group participants rather than the information possessed by only one or a few members is extensively explored in the research on group dynamics. This tendency is called *shared information bias*. An experiment carried out by Garold Stasser is a good illustration of this phenomenon. A group of eight people were asked to rate the performance of 32 psychology students. All members of the group received two relevant sources of information describing the students (e.g., grades) while two members of the group received two additional sources of information (e.g., degree of class participation) and one member received another source of information. Although the group as a whole then had five relevant sources of information available, their assessments were based almost exclusively on the information everyone shared (Surowiecki 2005). If making a correct decision hinges on examining the unshared information held by group members, the shared information bias can have adverse effects on decision quality.

Shared information bias can be reduced by increasing the diversity of opinion within a group (Smith 2008). Experiments show that heterogeneous groups, in which group members preferred different candidates prior to the discussion, were more likely to repeat unshared information than homogeneous groups.

To reap the benefits of diversity, however, it is not enough for all relevant expertise to be represented. The composition should also ensure a balanced representation of different expertise. This is important because the majority has a very strong influence due both to *informational conformity* (resulting when there are more arguments supporting the favored position) and to *normative conformity* (as people are less likely to want to be seen as disagreeing with the majority opinion).

Not agreeing with the group majority can be very uncomfortable. Berns et al. (2008) found in their experiments that this can cause brain activity associated with pain, fear, and stress. However, when minorities have an ally, disagreeing with the majority seems to be easier. Asch (1955) and (1956) found in his experiments that minority conformity rates were cut by one fourth when he added an ally compared to the experiments where the minority faced a unanimous majority.

Interpretation of Evaluation Criteria

Forscher et al. (2019) contend that if reviewers do not agree on their criteria, this will result in arbitrary and unreliable judgements. They posit that shared criteria enable and facilitate the consistency of reviews across reviewers as they provide a common and stable frame of reference. However, empirical evidence suggests that review criteria are not necessarily shared across reviewers. A survey of all humanities scholars in Switzerland provides evidence against the assumption that there is a consensus on evaluation criteria (Hugh and Ochsner 2022).

Other research in the field shows that reviewers weigh differently the various evaluation criteria they are assigned. For example, Abdoul et al. (2012) report that “Some reviewers gave the same weight to each assessment criterion, whereas others considered originality to be the most important criterion (12/34), followed by methodology (10/34) and feasibility (4/34). Conceivably, this variability might adversely affect the reliability of the review process.” A review of three American research funding organizations—National Institutes of Health, National Science Foundation and the National Endowment for the Humanities—showed that “unwritten or informal criteria were used by panels at all three agencies” (Government Accounting Office 1994, p. 83).

5. Disagreement

To analyze and measure disagreement in peer review, we now introduce a simple model of information processing with Bayesian updating. Bayesian updating is the process of revising assessments on the basis of additional information.

Suppose that there be $j = 1,2,\ldots,J$ projects in a single field. The projects could correspond to grant applications to a funding agency or article submissions to a scientific journal. Suppose that the merit of each project, denoted as θ_j, is distributed according to a normal distribution with mean μ and standard deviation σ_θ, representing the prior distribution in the population of applicants.

Suppose then that there are $i = 1,2,\ldots,I$ experts in the field. Each expert i receives a signal $x_i = \theta_j + \sigma_\varepsilon \varepsilon_i$ drawn from a normal distribution centered around the merit of the project θ_j,

with standard deviation σ_ε. For simplicity, assume that this standard deviation σ_ε is identical across projects and captures the noise level in each expert's evaluation of the merit of projects.

Focusing on a single expert, we can then compute the expert's conditional expectation of the merit θ_j conditional on receiving signal x_i. This conditional expectation is the best prediction of the project's merit and thus is an important benchmark for our analysis. In an ideal setting, the expert with perfectly aligned incentives would score the project by truthfully reporting this conditional expectation. Given that both the prior and the signal are normally distributed, the expert's conditional expectation, or score, is a linear combination of the prior mean μ and the signal x_i. The weight assigned to the signal increases in the prior variability of the merit of projects, σ_θ, and decreases in the signal noise, σ_ε. Intuitively, when the signal is very noisy the expert's score is equal to the prior μ and when the signal is without noise the expert's score is equal to actual merit θ_j.

The variance in the scores across experts evaluating any given project, thus holding fixed its merit θ_j, is equal to

$$Var(E[\theta|x_i]|\theta_j) = \frac{{\sigma_\theta}^4 {\sigma_\varepsilon}^2}{({\sigma_\theta}^2 + {\sigma_\varepsilon}^2)^2}.$$

Note that this variance is at its lowest when the noise in the signal is either zero or becomes arbitrarily large. Intuitively, when there is no noise, the score is perfect and equal to the project's merit θ_j, and so it is a constant conditional on θ_j. Similarly, when the noise is very high, the signal received contains no information and thus the score is equal to the prior μ, again a constant. Given that the variability of scores is non-monotonic in σ_ε, it is *not* a good indicator of the information content, $1/\sigma_\varepsilon$, of expert scores.

The correlation across the scores given by different experts, instead, is a much better measure of agreement. For the purpose of illustration consider the case with two experts, i and i'. Conditional on a project's merit θ_j, the correlation

$$Corr(E[\theta|x_i], E[\theta|x_{i'}]|\theta_j) = \frac{1}{{\sigma_\varepsilon}^2}(\theta - \mu)^2$$

decreases as the noise in the field increases. In addition, it does not depend on the variance of the merit of the projects in the field. The unconditional correlation of scores is instead

$$Corr(E[\theta|x_i], E[\theta|x_{i'}]) = \frac{{\sigma_\theta}^2}{{\sigma_\theta}^2 + {\sigma_\varepsilon}^2},$$

which decreases in σ_ε and increases in σ_θ. Intuitively, if the experts are perfectly informed and the signal contains no noise, the scores will be perfectly correlated. As the noise in the merit of projects increases, the scores given by reviewers become less positively correlated.

Measuring Disagreement

The correlation we just derived is compatible with the intraclass correlation ICC, a commonly used measure of the agreement of ratings given by different observers. At least since Zuckerman and Merton (1971), it has been widely believed that disagreement is more

pervasive in social sciences and humanities relative to physical sciences. While natural sciences such as physics and biology build on a core of widely agreed knowledge, this is not always the case in the social sciences.

As forcefully argued by Cole (1992), however, it is important to distinguish between core knowledge and frontier knowledge. Scientific knowledge is defined to be at the core of a field once the scientific community in the field reaches substantial consensus. Frontier knowledge instead comprises the latest scientific contributions. To put it simply, core knowledge forms the major part of the key textbooks in the field, while frontier knowledge is to be found in the latest journal publications. The structure of the two is also radically different. Core knowledge is ordered and has been refined over the years. Frontier knowledge is less structured and more chaotic, leading to higher levels of disagreement.

Summarizing a wide range of empirical research, Cole (1992) concludes that, even though natural sciences show a higher degree of consensus than social sciences on core knowledge, it is not clear that a clear difference can be discerned at the research frontier. He believes that the building of scientific knowledge through the operation of the reward system and disagreement work in a roughly similar way across all fields.

Measuring the level of consensus is not a straightforward task, as recognized by Cole (1992) in his empirical analysis. A critical step is defining the unit of analysis. The unit of analysis may be a general field (i.e., physics) or a specific subfield (i.e., nuclear physics), for example. Subfields comprise varying amounts of core and frontier knowledge. There is also significant overlap across fields so that it becomes difficult to define precisely an area over which to measure consensus. Similarly, measurements of disagreement in peer review of scientific journals need to take into account the scope of the journal, which is not always clear.

In an important study, Cicchetti (1991) investigates the level of agreement across reviewers depending on the breadth of the field covered by academic journals and grant evaluation panels. The generality of the field affects the level of disagreement. In general and diffuse fields (which attract a broader set of submissions and require more cross-disciplinarity), he finds more agreement on rejection than acceptance decisions. Instead, in more specific and narrower fields (which attract submissions more focused on a shared methodology), peer reviewers tend to agree more on acceptance than on rejection.

Overall, the large empirical literature on disagreement in grant peer review displays relatively low inter-rater reliability. While some researchers have suggested that inter-rater reliability for two reviewers on a single submission should be about 0.8–0.9 (Marsh et al. 2008, p. 162), actual agreement rates are considerably lower. For peer review procedures in funding organizations, reliability measures above 0.40 have rarely been demonstrated (Cicchetti 1991). More recently, Jayasinghe et al. (2003) reported a single-rater reliability correlation of just 0.21 for the humanities and social sciences, and an even lower correlation of 0.19 for the sciences. Fogelholm et al. found an inter-rater reliability of around 0.23 for medical research (Fogelholm et al. 2012). Similarly, Reinhart (2009) found an intraclass correlation coefficient for medicine of 0.20 and a higher rate for biology of 0.45.

As exemplified in the above studies, disagreement rates vary between fields of research, and there is evidence that increasing diversity of discipline of peer reviewers reduces rating consistency. Lobb et al. (2013) found that the level of agreement among experts from different disciplines was considerably lower than that among experts from the same disciplines, with an intraclass correlation coefficient of 0.66 for researchers from the same discipline and − 0.04 for adjudicators from different disciplines.

As grant peer review usually includes a panel stage after the initial round of individual reviews, studies have also been carried out comparing agreement rates between groups of reviewers, assembled in different review panels. The results indicate the presence of considerable disagreement there too.

In a seminar investigation, Cole et al. (1981) conducted a study of a real evaluation panel in NSF and a surrogate panel that was put together to resemble the original panel. The surrogate panel supported the real panel's decisions to fund / reject applications in 75% of cases, leading the authors to conclude that “the fate of a particular application is roughly half determined by the characteristics of the proposal and the principal investigator, and about half by apparently random elements which might be characterized as `the luck of the reviewer draw’.” A later study conducted by Hodgson (1997) yielded very similar results; two comparable panels that considered the same set of applications agreed to fund / reject applications in 73% of cases. A study conducted by Mayo et al. (2006) provided similar conclusions.

Is Disagreement Natural? Even Desirable?

While the majority of studies of inter-rater disagreement and its origins view disagreement and its root causes as problematic, others believe that “too much agreement is in fact a sign that the review process is not working well, that reviewers are not properly selected for diversity, and that some are redundant” (Bailar and Patterson 1985). Along the same lines, Reinhart (2009) points out that “complete agreement among reviewers would render it pointless to consider more than one review per application.” Lee (2013) qualifies as “questionable”, the assumption that disagreement among reviewers is not normatively appropriate or desirable, and asserts that “a natural direction for future research includes articulating and assessing alternative normative models that acknowledge reviewer partiality, with a focus on the epistemic and cultural bases for reviewer disagreement.”

According to these authors, inter-rater disagreement reflects normatively appropriate disagreements about what constitutes good research. Editors and grant program officers may seek reviewers who can evaluate different aspects of a submission according to their expertise (Bailar 1991), and high inter-rater reliability can not be expected in cases where different fields of expertise have varying perceptions of quality and focus on different parts of the application. In line with this, Lamont (2009) and Mallard et al. (2009) argue that when evaluative cultures are specific to disciplines, such differences should not necessarily be understood as a form of problematic bias.

Research on the social dynamics of groups concur with the view that diversity of opinion is not problematic, and can even be positive with respect to ensuring the quality of decisions.

Studies show that heterogeneous groups with complementary skills make better decisions than homogeneous groups (Levi 2007). Kahneman et al. (2021, p. 272) find that “the validity of pooled judgements increases faster when the judgements are uncorrelated with one another” and that the average of such an uncorrelated group paradoxically will be more accurate than the average of a unanimous one.

As discussed in the above, diversity can also serve to guard against negative group dynamics such as shared information bias (Smith 2008) and conformity pressure (Asch 1955, 1956).

6. Improving Reliability of Grant Review

While much work still remains to be done in terms of identifying measures that might serve to optimize the evaluation process, the literature on grant peer review and on social psychology suggests some key measures that might have an effect in ensuring a reliable process, minimizing the biases and process deficiencies discussed in the above. These include ensuring that a sufficient number of reviewers assess each application, supplying guidelines and training and ensuring structured assessments.

Sufficient Number of Reviewers

Research on grant peer review shows that "funding decisions will vary widely with the number of reviewers considering proposals that are closely scored" (Kaplan et al. 2008, p. 3). Graves et al. (2011) compare financing decisions made by panels consisting of 7, 9 and 11 members and finds that: “Reliability can be increased by using the most effective system of 11 panel members. This is probably worthwhile . . .” The challenge is that the use of a high number of experts results in high costs: “Use of additional reviewers per application may increase reliability and decision consistency, but adds to overall cost and burden” (Snell 2015).

Mayo et al. (2006) find through their statistical modeling that on average it is necessary for seven experts to evaluate each application in order to make robust, stable decisions. Marsh et al. (2008) find that at least six experts per application are needed to make decisions that are sufficiently robust. Snell (2015) concludes on the basis of statistical modeling that “Having too few reviewers (ie, N ≤ 3) results in excessive decision inconsistency . . . Five reviewers per application represents a practical trade-off, in terms of balancing increased decision consistency against incremental cost, as well as minimizing large random effects in decision outcomes and improving efficiency of the decision making process.”

The Working Group on Peer Review of the Advisory Committee to the Director of NIH (NIH 2008a) recommended increasing the number of experts evaluating each application in the NIH to four (NIH 2008b). Based on its review of over 30 research funding bodies, the ESF (2011) recommends that “the aim should be to provide at least three expert assessments before a final decision is made.”

While the literature on peer review seems to indicate that “the more the merrier” holds true, the social psychology literature on group dynamics however finds a tendency for groups to become less effective when size increases. Mullen and co-authors found in their meta review of eight US studies that the most effective working groups were of relatively small size—about four or five members. They also found that working in groups of about this size is more enjoyable to the members, in comparison with being in larger groups (Mullen et al. 1989). One reason that people are more motivated to put in more effort when groups are smaller is that each person plays a larger role in determining the outcome (Kameda et al. 1992).

Guidelines and Training

ESF (2011, p. 27) recommends that “before the tasks of both individual/remote reviewers and panel members begin, it is essential that their assignments are clearly described and communicated.” In general, reviewers demand clear guidelines from research funding bodies - a survey among reviewers used in application evaluation at nine different research funding bodies internationally showed that only 16% (42/258) considered that the research councils provided clear guidelines for their assignment. In total, 85% (220/258) had not received any training in assessing project applications and 64% (166/258) wanted this (Schroter et al. 2010).

There is some emerging evidence that training can increase the inter-rater reliability of reviewers. An online training video was found to increase the inter-rater reliability for both novice and experienced reviewers for the US National Institute of Health, with correlation scores rising from 0.61 to 0.89 following training (Sattler et al. 2015). Furthermore, the Canadian Institutes of Health Research used a training module on unconscious biases to address an identified bias against women in its grant system. Following its introduction female and male scientists had equal success rates (Guglielmi 2018).

The effect of training with respect to reducing unwanted variability in judgements has also been proven in a number of other fields (Kahneman et al. 2021). Training in the form of “learning by doing” has also proved effective in increasing inter-rater reliability. Studies on the issue find that differences in how reviewers translate their assessment onto a numerical scale is exacerbated if each assessor only makes ratings of one or a few submissions. In such cases, reviewers do not have a sufficient frame of reference for translating subjective impressions about the quality of a proposal onto the numerical scale that constitutes the peer review ratings. Jayasinghe et al. (2003, p. 298) find that: “The more proposals an assessor reviewed, the better the quality of the ratings they provided.” Jayasinghe et al. (2006, p. 594) conclude that “assessors who review a large number of proposals give lower, more reliable ratings than those who review a few proposals.”

That skills matter a lot is confirmed in studies of inter-rater reliability in other fields. In medicine, for example, a study of pneumonia diagnosis by radiologists found that “variation in skill can explain 44% of the variation in diagnostic decisions” (Kahneman et al 2021, p. 275).

Structured Assessments

Research on peer review shows that discussing criteria one by one provides a greater degree of equal treatment of applications and more consistent assessments of quality. Differences in the foci of committee discussions will introduce variation in the treatment of applications. For one application, a concern with methodology might dominate the discussion. For the next, a question about a particular publication may draw the committee's attention. By discussing a set of predefined criteria one by one, you ensure that the same things are emphasized in the discussion of each application within the panel and also a greater degree of equal treatment of applications across panels (Obrecht et al. 2007, Thornley et al. 2002, and Wood and Wessely 2003).

Kahneman et al. (2021) suggest a method designed to reduce noise, the so-called "mediating assessment protocol." The idea is to divide a complex assessment into mediating assessments—much like the different review criteria used in grant peer review. Decision makers should then evaluate the mediating assessments separately and explicitly. On each assessment, participants should make their judgements individually before discussing them one by one and making a final decision. Kahneman (2011, p. 224-225) argues that this is also in line with research in psychology which finds that "humans are incorrigibly inconsistent in making summary judgements of complex information. . . . Research suggests a surprising conclusion: to maximize predictive accuracy, final decisions should be left to formulas."[7]

While some funding agencies calculate the main score for each application based on the individual scores agreed for each criterion, others employ a practice whereby the overall score is set by the experts, based on the individual review criteria, but allowing for reviewer discretion in how to weigh the different criteria. Given the findings in research on peer review on how variability in the weights assigned to different criteria by different experts might result in reduced reliability of assessments, serious consideration should be given to experiment with ways to determine the final grant scores on the basis of the scores for each review criteria, rather in a discretionary manner by experts.

Failure to Fund Novel, Risky or Interdisciplinary Research

That peers disagree in their assessment of grant proposals is not surprising given the discussion above regarding the uncertainty associated with the scientific endeavor. This uncertainty is particularly pronounced in the context of research grant evaluation, as the evaluation is carried out ex ante and there is no guarantee that projects will achieve the objectives promised. As discussed, uncertainty associated with what Kuhn terms revolutionary science is significantly higher than that associated with what he terms normal science. How can revolutionary research be identified and supported effectively? Studies suggest that reviewer disagreement is higher for research grant proposals that attempt more radical shifts in the knowledge frontier compared to more incremental work. They recommend taking high disagreement among reviewers as an indicator of work with high risk

[7] McCauley (1991) finds that "Recent reviews draw on nearly 100 studies involving many different kinds of predictions . . . The results are strikingly consistent in showing that the formula does at least as well as the expert judge."

but potentially high reward and target such proposals for funding (Giraudeau et al. 2011, Linton 2016).

As the outcome of panel discussions are normally determined either by panel members agreeing on a consensus score based on the individual reviews of the different panel members, or by computing an average of the individual reviews, proposals where there is substantial disagreement are disadvantaged in the competition.

The literature on the subject points to a number of possibilities for adapting the review process to accommodate more risky/radically interdisciplinary proposals:

- Removing outlier best and worst scores from consideration thereby allowing proposals that cannot achieve consensus to go forward (Liaw et al. 2017).

- Not requiring consensus decisions in panel review, for example by allowing each panel member one “gold” vote, which allows them to protect a proposal they think is highly promising regardless of negative views of other panel members. Alternatively, issue reviewers with several gold votes, but still a limited supply. This forces reviewers to think carefully about how to allocate their stars across projects (Kolev et al. 2019).

- Asking reviewers to score peer review criteria separately and then curate a portfolio of accepted submissions that include some percentage of submissions that score highly for qualities typically underweighted by reviewers such as groundbreaking potential (OECD 2021).

- Identifying proposals with large disagreement between evaluators and funding these, based on the assumption that risky proposals will be subject to more controversy than safer proposals. For example, the NIH uses such an "out of order funding" approach where a number of applications for innovative research are chosen for funding despite receiving lower scores than other funded research based purely on the peer review process (Linder and Nakamura 2015).

- Asking evaluators to use less fine-grained scoring systems, which leads to more ties between competing submissions and allows funders to use funding priorities such as prioritization of particularly innovative research to serve as tiebreakers (Lee 2015).

- Introducing a lottery element in the review process. The idea is to first rely on peer reviewers to separate applications that are worthy of funding from those that are not, and then select grant winners among those deemed worthy of funding based on a lottery (Fang and Casadevall 2016). Advocates argue that a lottery can reduce some of the bias associated with more risky/radically interdisciplinary proposals, as it would be possible for proposals subject to reviewer disagreement to make it into the fundable lottery pool, but less likely that they would make it to the top of a ranked list decided by reviewers. Especially in highly competitive calls, the top ranked positions are normally reserved for those proposals where there is universal agreement about the merit. Given the traction this idea is gaining in policy circles, it is an interesting challenge for mechanism designers to investigate under what condition a lottery-

based grant allocation can be optimal.

References

Abdoul H., C. Perrey, P. Amiel, F. Tubach, S. Gottot, et al. 2012. Peer Review of Grant Applications: Criteria Used and Qualitative Study of Reviewer Practices. *PLoS ONE* 7(9): e46054.

Albright, Alice, Michael Kremer, Ruth Levine, and Center for Global Development,(editors). 2005. *Making Markets for Vaccines: Ideas to Action: Report of the Center for Global Development Working Group*. Washington, D.C.: Center for Global Development.

Arensbergen, Pleun van, Inge van der Weijde and Peter van den Besselaar. 2014. "The Selection of Talent as a Group Process. A Literature Review on the Social Dynamics of Decision Making in Grant Panels." *Research Evaluation* 23(4): 298–311.

Arrow, Kenneth J. 1951. *Social Choice and Individual Values*. Yale University Press.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In: *The Rate and Direction of Inventive Activity* (edited by Richard Nelson), 609–626. Princeton University Press.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53(5): 941–973.

Asch, Solomon E. 1955. "Opinions and Social Pressure." *Scientific American* 193: 31–35.

Asch, Solomon E. 1956. "Studies of Independence and Conformity: I. A Minority of One against a Unanimous Majority." *Psychological Monographs: General and Applied* 70(9): 1–70.

Azoulay, Pierre, and Danielle Li. 2020. "Scientific Grant Funding." NBER Working Paper 26889.

Bailar, John C. 1991. "Reliability, Fairness, Objectivity and Other Inappropriate Goals in Peer Review." *Behavioral and Brain Sciences* 14: 137–38.

Bailar J. C. and K. Patterson 1985. "Journal Peer Review: The Need for a Research Agenda." *New England Journal of Medicine* 312: 654–657

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics* 107, no. 3: 797–817.

Berns, Gregory S., Jonathan Chappelow, Caroline F. Zink, Giuseppe Pagnoni, Megan E. Martin-Skurski, and Jim Richards. 2005. 'Neurobiological Correlates of Social Conformity and Independence During Mental Rotation', Biol Psychiatry, 58: 245–253.

Biagioli, Mario. 1990. "Galileo's System of Patronage." *History of Science* 28: 1–79.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. “A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades.” *Journal of Political Economy* 100(5): 992–1026.

Blank, Rebecca M., 1991. “The Effects of Double-blind versus Single-blind Reviewing: Experimental Evidence from the American Economic Review. *American Economic Review* 81(5): 1041–1067.

Bloom, David E., Daniel Cadarette, Maddalena Ferranna, Randall N. Hyer, and Daniel L. Tortorice. 2021. “How New Models of Vaccine Development for COVID-19 Have Helped Address an Epic Public Health Crisis.” *Health Affairs* 40(3): 410–418.

Bornmann, Lutz. 2011. “Scientific Peer Review.” *Annual Review of Information Science and Technology* 45(1): 197–245.

Boudreau, Kevin J., Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl. 2016. “Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science.” *Management Science* 62(10): 2765–2783.

CDC. 2022. “Understanding MRNA COVID-19 Vaccines.” Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/mrna.html.

Cicchetti, Domenic V. 1991. “The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation.” *Behavioral and Brain Sciences* 14(1), 119–135.

Cobb, Matthew. 2015. “Who Discovered Messenger RNA?” *Current Biology* 25(13): R526–32.

Cole, Stephen. 1992. *Making Science: Between Nature and Society*. Harvard University Press.

Cole, Stephen, Jonathan Cole, and Gary Simon. 1981. “Chance and Consensus in Peer Review.” *Science* 214: 881–886.

Crosland, Maurice, and Antonio Gálvez. 1989. “The Emergence of Research Grants within the Prize System of the French Academy of Sciences, 1795-1914.” *Social Studies of Science* 19(1): 71–100.

Dasgupta, Partha, and Paul A. David. 1994. “Toward a New Economics of Science.” *Research Policy* 23(5): 487–521.

David, Paul A. 1998. “Common Agency Contracting and the Emergence of ‘Open Science’ Institutions.” *American Economic Review* 88(2): 15–21.

David, Paul A. 2008. “The Historical Origins of ‘Open Science’: An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution.” *Capitalism and Society* 3(2): 5.

Dolgin, Elie. 2021. “The Tangled History of mRNA Vaccines.” *Nature* 597(7876): 318–324.

European Science Foundation. 2011. *European Peer Review Guide. Integrating Policies and Practices into Coherent Procedures*. Strasbourg.

Fang F. C., Bowen A., A. Casadevall. 2016. “Research: NIH Peer Review Percentile Scores are Poorly Predictive of Grant Productivity.” *eLife* 5:e13323.

Feynman, Richard. 1998. “The Uncertainty of Science.” Lecture 1 in: *The Meaning of It All.* Addison-Wesley.

Fogelholm, Mikael, Saara Leppinen, Anssi Auvinen, Jani Raitanen, Anu Nuutinen, and Kalervo Väänänen. 2012. “Panel Discussion Does Not Improve Reliability of Peer Review for Medical Research Grant Proposals.” *Journal of Clinical Epidemiology* 65, no. 1: 47–52.

Forscher, Patrick S., Markus Brauer, Flavio Azevedo, William T. L. Cox, and Patricia G. Devine. 2019. “How Many Reviewers Are Required to Obtain Reliable Evaluations of NIH R01 Grant Proposals?” *PsyArXiv*. April 9.

Franzoni, Chiara, Paula Stephan, and Reinhilde Veugelers. 2021. “Founding Risky Research.” NBER Working Paper 28905.

Garfagnini, Umberto, Marco Ottaviani, and Peter N. Sørensen. 2014. “Accept or Reject? An Organizational Perspective.” *International Journal of Industrial Organization* 34: 66-74.

Government Accounting Office. 1994. *Peer Review: Reforms Needed to Ensure Fairness in Federal Agency Grant Selection*. Washington, DC.

Gallo, Stephen A., Joanne H. Sullivan, and Scott R. Glisson. 2016. “The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications.” *PLoS One* 11:e0165147.

Geuna, Aldo, and Ben R. Martin. 2003. “University Research Evaluation and Funding: An International Comparison.” *Minerva* 41: 277–304.

Giraudeau, Bruno, Clémence Leyrat, Amélie Le Gouge, Julie Léger, and Agnès Caille. 2011. “Peer Review of Grant Applications: A Simple Method to Identify Proposals with Discordant Reviews.” *PLoS ONE* 6(11): e27557.

Gleick, James. 1992. *Genius: Richard Feynman and Modern Physics*. Abacus London.

Goldin, Claudia, and Cecilia Rouse. 2000. “Orchestrating Impartiality: The Impact of “Blind’ Auditions” on Female Musicians.” *American Economic Review* 90(4): 715–741.

Graves N., A. G. Barnett, and P. Clarke. 2011. “Funding Grant Proposals for Scientific Research: Retrospective Analysis of Scores by Members of Grant Review Panel.” *BMJ* 343: d4797.

Guglielmi, Giorgia. 2018. “Gender Bias Goes Away when Grant Reviewers focus on the Science.” *Nature* 554: 14-16.

Guthrie, S., I. Ghiga, and S. Wooding. 2018. “What Do We Know about Grant Peer Review in the Health Sciences?*” F1000Research* 6.

Hanson, Robin. 1998. “Patterns of Patronage: Why Grants Won Over Prizes in Science.” Unpublished Paper.

Herbert, Danielle L., Adrian G. Barnett, Philip Clarke, and Nicholas Graves. 2013. “On the Time Spent Preparing Grant Proposals: An Observational Study of Australian Researchers.” *BMJ Open* 3:e002800.

Hemel, Daniel, and Lisa Ouellette. 2013. “Beyond the Patents-Prizes Debate.” *Texas Law Review* 92(2): 303–82.

Hodgson C. 1997. “How Reliable is Peer Review? A Comparison of Operating Grant Proposals Simultaneously Submitted to Two Similar Peer Review Systems.” *Journal of Clinical Epidemiology* 50: 1189-1195.

Holmstrom, Bengt, and Paul Milgrom. 1994. “The Firm as an Incentive System.” *American Economic Review* 84(4): 972–91.

Hugh, S. E., and M. Ochsner. 2022. “Do Peers Share the Same Criteria for Assessing Grant Applications?” *Research Evaluation* 31(1): 104-117.

James, Estelle. 1990. “Decision Processes and Priorities in Higher Education.” In: *The Economics of American Universities: Management, Operations and Fiscal Environment* (Edited by Stephen A. Hoenack and Eileen L. Collins). Albany, State University of New York Press.

Janis, Irving L., 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin Company, Boston.

Jayasinghe U. W., Marsh H. W., and N. Bond. 2003. “A Multilevel Cross-Classified Modelling Approach to Peer Review of Grant Proposals: The Effects of Assessor and Researcher Attributes on Assessor Ratings.” *Journal of the Royal Statistical Society Series A* 166(3): 279–300.

Jayasinghe, U.W., Marsh, H. W., and N. Bond. 2006. “A New Reader Trial Approach to Peer Review in Funding Research Grants: An Australian Experiment.” *Scientometrics* 69: 160-168

Jeffs, Lloyd B., Lorne R. Palmer, Ellen G. Ambegia, Cory Giesbrecht, Shannon Ewanick, and Ian MacLachlan. 2005. “A Scalable, Extrusion-Free Method for Efficient Liposomal Encapsulation of Plasmid DNA.” *Pharmaceutical Research* 22(3): 362–72.

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. London: Allen Lane.

Kalil, Thomas. 2006. “Prizes for Technological Innovation.” Discussion Paper. The Hamilton Project. Brookings Institution.

Kameda, Tatsuya, Mark F. Stasson, James H. Davis, Craig D. Parks, and Suzi K. Zimmerman. 1992. “Social Dilemmas, Subgroups, and Motivation Loss in Task-Oriented Groups: In Search of an ‘Optimal’ Team Size in Division of Work.” *Social Psychology Quarterly* 55(1): 47–56.

Kaplan D., Lacetera N., and C. Kaplan. 2008. “Sample Size and Precision in NIH Peer Review.” *PLoS ONE* 3(7): e2761.

Ke, Qing, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. 2015. “Defining and Identifying Sleeping Beauties in Science.” *Proceedings of the National Academy of Sciences* 112(24): 7426–7431.

Kirchdoerfer, Robert N., Christopher A. Cottrell, Nianshuang Wang, Jesper Pallesen, Hadi M. Yassine, Hannah L. Turner, Kizzmekia S. Corbett, Barney S. Graham, Jason S. McLellan, and Andrew B. Ward. 2016. “Pre-Fusion Structure of a Human Coronavirus Spike Protein.” *Nature* 531(7592): 118–21.

Kolata, Gina, and Benjamin Mueller. 2022. “Halting Progress and Happy Accidents: How mRNA Vaccines Were Made.” *New York Times*, 15 January. https://www.nytimes.com/2022/01/15/health/mrna-vaccine.html.

Kolev, Julian, Pierre Azoulay, Yuly Fuentes-Medel, and Fiona Murray. 2019. “Expert Evaluation in Innovation: The Role of Distance and Consensus in Project Selection.” *Working Paper*, Southern Methodist University.

Kortum, Samuel, and Lerner, Josh. 2000. “Assessing the Contribution of Venture Capital to Innovation”. *RAND Journal of Economics* 31(4): 674–692

Kremer, Michael, Jonathan Levin, and Christopher M. Snyder. 2020. “Advance Market Commitments: Insights from Theory and Experience.” *AEA Papers and Proceedings* 110: 269–73.

Kuhn, Thomas. 1996. *The Structure of Scientific Revolutions*. Third, Chicago: University of Chicago Press.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment.* Cambridge, MA: Harvard University Press.

Langfeldt, Liv, and Svein Erik Moen. 2013. “Regionale forskningsmidler i Helse Sør-Øst RHF: Evaluering av ressursbruk og habilitet i tildelingsprosessen.”

Langfeldt, Liv. 2002. Decision-Making in Expert Panels Evaluating Research. Constraints, Processes and Bias.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. “Bias in Peer Review.” *Journal of the American Society for Information Science and Technology* 64(1): 2–17.

Lee, C. J. 2015. “Commensuration Bias in Peer Review”. *Philosophy of Science*, *82*(5), 1272–1283.

Levi, D. 2007. *Group Dynamics for Teams*. London, Sage. (2014, 4th edition)

Li, Danielle, 2017. “Expertise versus Bias in Evaluation: Evidence from the NIH.” *American Economic Journal: Applied Economics* 9(2): 60–92.

Liaw, L., J. E. Freedman, L. B. Becker, N. N. Mehta and L. Liscum. 2017. “Peer Review Practices for Evaluating Biomedical Research Grants: A Scientific Statement from the American Heart Association.” *Circulation Research* 121(4): e9-e19.

Lindner, M. D. and R. K. Nakamura. 2015. “Examining the Predictive Validity of NIH peer Review Scores”. *PLoS One*, *10*(6): e0126938.

Linton, Jonathan D. 2016. “Improving the Peer Review Process: Capturing More Information and Enabling High-Risk/High-Return Research.” *Research Policy* 45(9): 1936–1938.

Lobb, Rebecca, Lisa Petermann, Elizabeth Manafo, Deb Keen, and Jon Kerner. 2013. “Networking and Knowledge Exchange to Promote the Formation of Transdisciplinary Coalitions and Levels of Agreement Among Transdisciplinary Peer Reviewers.” *Journal of Public Health Management and Practice* 19(1): E9.

Lotka, Alfred J. 1926. “The Frequency Distribution of Scientific Productivity.” *Journal of the Washington Academy of Sciences*, 16(12): 317–323.

Mallard, Grégoire, Michèle Lamont, and Joshua Guetzkow. 2009. “Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review.” *Science, Technology, and Human Values* 34: 573–606.

Marschak, Jacob, and Roy Radner. 1972. *Economic Theory of Teams*. New Haven: Yale University.

Marsh, Herbert W., U.W. Jayasinghe and N. Bond. 2008: “Improving the Peer Review Process for Grant Applications. Reliability, Validity, Bias and Generalizability.” *American Psychologist* 63/3: 160-8

Marsh, Herbert W., and Patricia Bazeley. 1999. “Multiple Evaluations of Grant Proposals by Independent Assessors: Confirmatory Factor Analysis Evaluations of Reliability, Validity, and Structure.” *Multivariate Behavioral Research* 34: 1–30.

Maurer, Stephen, and Suzanne Scotchmer. 2004. “Procuring Knowledge.” *Intellectual Property and Entrepreneurship* 15: 1–31.

Mayo, N. E., Goldberg M. S., Klein M. B., Miller S., Platt R. W., and J. Richie. 2006. “Peering at Peer Review Revealed High Degree of Chance Associated with Funding of Grant Applications.” *Journal of Clinical Epidemiology* 8: 842–848.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüss, Michael Razen, and Utz Weitzel et al. 2021. “Non-Standard Errors.” SSRN Scholarly Paper.

Merton, Robert K. 1942. “A Note on Science and Democracy.” *Journal of Legal and Political Sociology* 1(1-2): 115–26.

Merton, Robert K. 1973. *“*The Sociology of Science: Theoretical and Empirical Investigations”. University of Chicago Press.

Mullen, Brian, Cynthia Symons, Li-tze Hu, and Eduardo Salas. 1989. “Group Size, Leadership Behavior, and Subordinate Satisfaction.” *Journal of General Psychology* 116: 155–170.

Name-Correa, Alvaro J., and Huseyin Yildirim. 2019. “Social Pressure, Transparency, and Voting in Committees.” *Journal of Economic Theory* 184: 104943.

National Science Foundation. 2022. “About NSF: Overview.” https://www.nsf.gov/about/.

Nature.com. 2021. “The Tangled History of mRNA Vaccines.” https://www.nature.com/articles/d41586-021-02483-w

NIH. 2008a. *National Institutes of Health 2007-2008 Peer Review Self-Study*. http://enhancing-peer-review.nih.gov/meetings/NIHPeerReviewReportFINALDRAFT.pdf

NIH. 2008b. *Enhancing Peer Review: A Self-Study by the NIH in Partnership with the Scientific Community to Strengthen Peer Review in Changing Times*. http://enhancing-peer-review.nih.gov/meetings/EnhancingPeerReviewACD2-21-08.pdf

Nunnari, Salvatore, Marco Ottaviani, and Debrah Meloso. 2018. “Looking into Crystal Balls: A Laboratory Experiment on Reputational Cheap Talk.” CEPR Discussion Paper DP1323.

Obrecht M., Tibelius K and G. D'Aloisio. 2007. Examining the Value Added by Committee Discussion in the Review of Applications for Research Awards. *Research Evaluation* 16: 79–91.

OECD. 2021. “Effective Policies to Foster High-Risk/High-Reward Research.” *OECD Science, Technology and Industry Policy Papers*, No. 112.

Ottaviani, Marco, and Peter Norman Sørensen. 2001. “Information Aggregation in Debate: Who Should Speak First?” *Journal of Public Economics* 81(3): 393–421.

Ottaviani, Marco, and Peter Norman Sørensen. 2006. “Reputational Cheap Talk.” *RAND Journal of Economics* 37(1): 155–75.

Ottaviani, Marco. 2020. “Grantmaking.” CEPR Discussion Paper 15389.

Pallesen, Jesper, Nianshuang Wang, Kizzmekia S. Corbett, Daniel Wrapp, Robert N. Kirchdoerfer, Hannah L. Turner, Christopher A. Cottrell, et al. 2017. “Immunogenicity and Structures of a Rationally Designed Prefusion MERS-CoV Spike Antigen.” *Proceedings of the National Academy of Sciences* 114(35): E7348–57.

Pardi, Norbert, Michael J Hogan, and Drew Weissman. 2020. “Recent Advances in MRNA Vaccine Technology.” *Current Opinion in Immunology*, Vaccines: Special Section on Evolutionary and Systems Immunology, 65: 14–20.

Pearce, Richard M. 1912. “Chance and the Prepared Mind.” *Science* 35(912): 941–956.

Pfizer. 2020. “Pfizer and BioNTech Announce an Agreement with U.S. Government for up to 600 Million Doses of MRNA-Based Vaccine Candidate Against SARS-CoV-2.” https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-announce-agreement-us-government-600.

Pigou, Arthus C. 1920. *The Economics of Welfare*. London: Macmillan.

Reale, Emanuela, and Antonio Zinilli. 2019. “Evaluation for the Allocation of University Research Project Funding: Can Rules improve the Peer Review?” *Research Evaluation* 26(3): 190–198.

Reinhart, Martin. 2009. “Peer review of grant applications in biology and medicine. Reliability, fairness, and validity.” *Scientometrics* 81: 789-809.

Rietzschel, Eric F., Bernard A. Nijstad, and Wolfgang Stroebe. 2019. “Why Great Ideas Are Often Overlooked: A Review and Theoretical Analysis of Research on Idea Evaluation and Selection.” In: *The Oxford Handbook of Group Creativity and Innovation* (edited by Paul B. Paulus and Bernard A. Nijstad). Oxford University Press.

Rothberg, Emma. 2021. “Kati Karikó.” National Women’s History Museum. https://www.womenshistory.org/education-resources/biographies/kati-kariko

Rovelli, Carlo. 2012. “Science Is Not about Certainty: A Philosophy of Physics.” Edge. https://www.edge.org/conversation/carlo_rovelli-science-is-not-about-certainty-a-philosophy-of-physics

Roy, Rustum. 1985. “Funding Science: The Real Defects of Peer Review and an Alternative to It.” *Science, Technology, and Human Values* 10(3): 73-81.

Sattler, David N., Patrick E. McKnight, Linda Naney, and Randy Mathis. 2015. 'Grant Peer Review: Improving Inter-Rater Reliability with Training', *PLoS One* 10: e0130450.

Scotchmer, Suzanne. 2004. *Innovation and Incentives*. Cambridge, MA: MIT Press.

Schroter S., Groves T., and L. Højgaard L. 2010. Surveys of Current Status in Biomedical Science Grant Review: Funding Organizations' and Grant Reviewers' Perspectives. *BMC Medicine* 8: 62.

Severin, Anna, Joao Martins, Rachel Heyard, François Delavy, Anne Jorstad, Matthias Egger. 2020. “Gender and Other Potential Biases in Peer Review: Cross-Sectional Analysis of 38 250 External Peer Review Reports.” *BMJ Open* 10:e035058.

Shapin, Steven. 1995. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press.

Smith, Christine M. 2008. “Adding Minority Status to a Source of Conflict: An Examination of Influence Processes and Product Quality in Dyads.*” European Journal of Social Psychology* 38(1): 75–83.

Snell R. R. 2015. Menage a Quoi? Optimal Number of Peer Reviewers. *PLoS ONE* 10(4): e0120838.

Sobel, Dava. 2007. *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. Bloomsbury.

Stephan, Paula. 2012. *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.

Surowiecki, James. 2005. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*. Random House: New York.

Taylor, Curtis R., and Huseyin Yildirim. 2011. “Subjective Performance and the Value of Blind Evaluation.” *Review of Economic Studies* 78(2): 762–94.

Thornley, R, M W Spence, M Taylor and J Magnan. 2002. New decision tool to evaluate award selection process. *Journal of Research Administration* 33(2): 49–56.

Wang, Jian, Reinhilde Veugelers, and Paula Stephan. 2017. “Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators.” *Research Policy* 46(8): 1416–1436.

Westfall, Richard S. 1985. “Science and Patronage: Galileo and the Telescope.” *Isis*, 76(1): 11–30.

Wood, Fiona, and Simon Wessely. 2003. “Peer Review of Grant Applications: A Systematic Review.” In: *Peer Review in Health Sciences, Second Edition*. London: BMJ Publishing Group.

World Health Organization. 2022. “WHO Coronavirus (COVID-19) Dashboard,”. https://covid19.who.int.

Wright, Brian D. 1983. “The Economics of Invention Incentives: Patents, Prizes, and Research Contracts.” *American Economic Review* 73(4): 691–707.

Zuckerman, Harriet, and Robert K. Merton. 1971. “Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System.” *Minerva* 9(1): 66–100.