DOI: 10.5644/PI2021.200.11

The limitations of scientific consensus

Key note address

Dražen Prelec[*]

It is a pleasure and honour to contribute to the conference on "Science System — a Factor of Stimulation or Limitation in Development," organized by the Academy of Sciences and Arts of Bosnia and Herzegovina. One cannot overstate the importance and timeliness of the conference topic, which raises at least two distinct questions: The first 'external' question is how to measure the impact of scientific research on society as a whole; the second, 'internal' question is how the scientific system should be organized to maximize the production of genuine scholarship. From the program, it is evident that both questions will be discussed during the course of the day.

I should acknowledge at the outset that my research is not in the study of science per se. I have worked on decision making, behavioural economics, and, what is relevant here, on the development of mechanisms for eliciting and aggregating expert judgment, which in terms of this conference, makes me a 'meta-expert' if not a 'domain expert.' Expert or specialized judgment is, of course, at the heart of science, contributing to tension between scientists and the general public, which is asked to accept the judgments of the scientific community without necessarily understanding the supporting reasons or evidence. The lay public is asked simply to trust science.

The focus of my talk will be on methods for reaching a consensus judgment in situations where experts disagree. Such disagreement arises routinely in every discipline and institution. It is endemic to legal decision making, artistic judgments, and in certain areas of medicine, notably psychiatry. Within natural science it has both an internal and external aspect. At the every-day

[*] Professor in the Sloan School of Management, the Department of Economics, and the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology. E-mail: dprelec@mit.edu.

level within the ordinary functioning of science, there are decisions about accepting articles for publication, funding grant proposals, or hiring and promoting individual faculty candidates. In such cases, equally qualified judges or panel members may reach different conclusions. How, then, should decisions be made? Should they be entrusted to a single 'dictator,' such as a journal editor or university president, or to a credentialed panel, as is typically the case in grant funding decisions? The challenge of reaching a consensus judgment also appears at the external level, when the discipline speaks to social problems and policy. The question then is what the discipline actually believes, for example about climate change or economic policy. Should that collective belief be computed by some kind of average of beliefs of card-holding professionals in the discipline?

We are touching here on an ancient problem of democratic voting, and the related questions of who is competent to vote on a particular question, and how votes should be combined to produce a decision. One could start with the ancient Greeks; I will only note two important more recent results, one mathematical and one empirical.

The mathematical result is Condorcet's 'jury theorem' (Condorcet, 1785). Condorcet considered the accuracy of majority opinion on binary question, such as whether a proposition is True or False. The question is presumed to have an objectively correct answer, which no single person necessarily knows. Condorcet proved that if each voting individual has a better than 50% chance of being correct, then majority opinion becomes an increasingly likely indicator of truth as the number of voters increases, and becomes infallible as that number approaches infinity. This was an early proof of the law of large numbers.

More than a century later, the English scholar Francis Galton provided a small but elegant empirical demonstration, which proved even more influential (Galton 1907). Galton conducted a judgmental experiment in the natural setting of a country fair, where animals were being bought and sold. He asked people there to judge the weight of a single ox, or, more precisely, the weight of meat produced by the animal after it had been slaughtered. Galton reasoned that the accuracy of the median estimate would constitute a fair test of democratic decision making:

> "According to the democratic principle of 'one vote one value' the middlemost [median] estimate expresses the *vox populi*, every other estimate being condemned as too low or too high by a majority of the voters..."

Galton did not screen for expertise, and the people providing estimates were a diverse lot, some experienced with assessing the value of animals offered for sale and others mere passers-by. The experiment could thus be presented as a test of the democratic one-person one-vote principle. The result of the experiment was remarkable and perhaps a disappointment to Galton who was politically conservative. The median estimate matched actual weight almost exactly.

Galton's paper become the stimulus to what is now known as the 'wisdom-of-the-crowd' philosophy, which holds that algorithmic aggregation of opinions is superior to, and should replace credentialed experts. The utopian claims associated with this philosophy are well summarized in the best seller "The Wisdom of Crowds," by the journalist James Surowiecki:

> "Large groups of people are smarter than an elite few, no matter how brilliant — better at solving problems, fostering innovation, coming to wise decisions, even predicting the future (Surowiecki, 2005)."

The wisdom-of-the-crowd movement draws strength from two trends - one technological and one social-political. The widespread adoption of electronic communication devices has made it much easier to tap vast numbers of dispersed individuals, and, if needed, pay them for offering their opinions. We see this in the proliferation of online ratings systems and freely provided evaluations; we can all vote on all kinds of different question without much effort. Galton's vox populi is now heard everywhere, all the time. This is accompanied by increased mistrust of expertise in general, and in scientific expertise in particular.

However, the technological advances in communicating opinions did nothing to solve an underlying theoretical problem: Although it is obviously true that a large group of people has more information than any single individual, it is not obviously true that unfiltered majority opinion will correctly reflect that information. Majority opinion can certainly produce the wrong result if the best information is concentrated amongst a small number of individuals, while the majority holds incorrect beliefs. This is true in science as well, as new scientific developments are often greeted with scepticism even within the scientific community itself.

Before considering alternatives to majority opinion as a definition of truth, we should note that our culture possesses another powerful mechanism for eliciting and combining opinions from many individuals, namely markets and market institutions. In markets, these opinions are not expressed explicitly, but only implicitly through trading activity. That is, by investing in a market, investors are declaring that they have some information or insight that is not reflected in the current market price, but will be reflected at a later time, yielding a profit as a result.

Regarded as an instrument of collective decision making, markets have three interesting features. First, markets are procedurally democratic: Everyone can participate; a person does not need a degree or pass a test to be able to ‘vote’ in a market. However, unlike voting, the market verdict is not algorithmically democratic because the price does not reflect the average opinion, but the average weighted by how much different market participants are willing to invest. Those with deeper pockets have more influence. In theory at least, the market instrument should be exquisitely sensitive to expertise, as individuals with superior information will invest relatively more and have more impact on the price. Third, markets solve what economists call the ‘incentive-compatibility problem,’ that is, they provide incentives for individuals to reveal their honest opinions about future prices. This is not the case with voting or rating, where people can vote or rate dishonestly without any penalty.

Given these advantages, one might ask whether markets could have a role to play in the allocation of resources to scientific projects, evaluation of grant proposals, or editorial decisions in journals? Indeed, recent work has shown that a certain type of ‘prediction market’ is able to predict the reproducibility of scientific findings (Dreber et al., 2015). The market participants are able to intuit, to some extent, whether a published result is fragile or not credible.

It remains to be seen whether markets can play a larger role in scientific decision making. The crucial limitation of markets is that market securities must be defined with respect to a verifiable, public metric or event. In the paper just mentioned, the verifiable event is the outcome of an attempt to reproduce an experimental finding. The result will be known: Either the experimental result reproduces or does not reproduce. But many, if not most interesting scientific questions are not readily verifiable. This is certainly true in social sciences, where, for example, the ‘truth’ of a Keynesian approach to macroeconomics will never be resolved to everyone’s satisfaction.

A similar problem appears if one were to attempt to use markets to make a grant funding decision. One can imagine a market where individuals would invest in proposals that are competing for funds. However, even if one could define a verifiable index for evaluating the ultimate performance of a research project, such an index could only be computed for proposals that were actually funded; unfunded proposals would not have determinate market value.

For these reasons some type of voting procedure remains the instrument of choice for reconciling opinions in science. As noted already, a major limitation of voting is its insensitivity to relative expertise. This problem cannot be solved by asking individuals to indicate their level of confidence and then weighting their answers by self-reported subjective confidence. People are simply not good in reporting confidence, and it is easy to imagine scenarios where experts are precisely the people who are relatively less confident (this would be the case when the information available to experts shatters confidence in a widely held consensus view).

In a recent paper my colleagues and I have proposed a solution to this problem, based on asking experts not only to give their personal judgment but also to predict the distribution of judgments that other experts, i.e., their 'peers,' will provide (Prelec, Seung and McCoy, 2017). The voting principle described in there is not the standard democratic principle of selecting the alternative that receives the most votes, i.e., the most popular alternative, but the different principle of selecting the alternative that receives the most votes relative to predictions. This is called the 'surprisingly popular' principle. The claims behind the theory are based on a mathematical model, but the main intuition can be conveyed through simple examples.[1]

As an elementary demonstration or 'proof-of-concept,' let us consider a simple factual question where most people provide the wrong answer. The following is an example of such a question: "Is the city of Philadelphia the capital of the US state of Pennsylvania?" Because Philadelphia is a familiar large city, and also plays an important role in US history, it seems like a natural candidate to be the state capital, and most people in the US believe that it is the state capital. However, the actual capital is the relatively unknown city of Harrisburg. Therefore, the Condorcet assumption fails and each voter has less than 50% chance of being correct on this question. The majority opinion is usually wrong, and, indeed, is guaranteed to be wrong in the large sample limit.

Here is how this problem can be fixed. The surprisingly popular principle does not endorse majority opinion, but instead compares the votes for each possible answer against predicted votes for that answer. In the Philadelphia case, even though only a minority of people (about 35%) provide the correct

[1] The surprisingly popular principle can also be used to provide incentives for honest judgments, in situations where honesty cannot be independently verified (Prelec, 2004).

answer “No”, this is more than the predicted votes for “No” which are about 25%. Therefore, the answer “No” will be declared as the correct answer according to the method, since 35% is greater than 25%.

It is worth pausing to consider why the predicted votes for “No” are fewer than the actual votes. This is because individuals who know that the correct answer is “No” also know that most people do not know the correct answer. Another way to express the intuition is that if Philadelphia were really the capital, then almost everyone would believe that to be the case, rather than only 65%. If the obvious answer is true, then it should be endorsed by almost everyone; if a significant minority disagrees with the obvious answer, then that answer is most likely false. Essentially, the surprisingly popular principle implements a handicapping system, where different answers are evaluated by comparing how much support they receive relative to the predicted handicap.

To provide a demonstration with more substantive interest, involving genuine domain expertise, we tested the method with art experts estimating the market value of modern artworks (Prelec, Seung & McCoy, 2017). From the standpoint of method, the challenge here is similar to one faced by a scientific panel of experts tasked to evaluate different research grant proposals. In both settings, the panel will be composed of people with different tastes and different levels of knowledge about individual artworks or proposal submitted for evaluation. In our study with modern art, we predicted and observed a specific type of failure of the democratic voting algorithm, namely, that it is too conservative. Even if a few experts do recognize an extraordinary talent, their opinions will generally be overcome by their colleagues who do not recognize it. It is important to understand this is not a problem of individual ignorance, but rather a problem of the voting method. An exceptional artwork has a better chance of being recognized by a single randomly selected panel member, than by the majority opinion of the entire panel. (One could say that these findings provide a kind of support for the practice of artistic patronage, where decisions are essentially delegated to a single aristocratic or wealthy individual).

It is plausible that the same problem arises within scientific panels or committees, because structurally the setting and the decision dilemmas are similar. Among the individuals who compose such panels, there will be differences in expertise in relation to specific proposals. Those who are relatively less familiar with a proposal are likely to default to a safe ‘average’ rating, which is the correct response at the personal level. This generates a bias against truly original proposals, whose merits are recognized by relatively few. In theory, the surprisingly popular voting principle will eliminate the bias, and provide more opportunities for new ideas to receive funding.

Such an experiment has not yet been done, and these proposals remain speculative. The challenges with implementing this method, and evaluating its performance relative to a traditional, majority opinion alternative, are not trivial. Nonetheless, we know enough about individual and collective decision-making biases to recommend a change in how things are done, in the hope of promoting ideas and projects that would not otherwise survive.

We have addressed so far the ‘internal’ question of improving resource allocation within science, but the same concerns and the potential remedies apply to the ‘external’ question of identifying and communicating the best answer to questions that society might pose to the scientific community. Mere scientific consensus is not a reliable indicator of the best answer in light of all scientific evidence, because evidence is not democratically distributed within the scientific community, which, like all human communities, is subject to conventions, fashion and prejudice.

Literature

Condorcet, J.-A.-N. de Caritat, marquis de. *Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale, 1785.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B.A., & Johannesson, M. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 2015, *112*(50), 15343-15347.

Galton, F. Vox populi (the wisdom of crowds). *Nature*, 1907, *75*(7), 450-451.

Prelec, D., Seung, H. S., & J. McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 2017, *541*(7638), 532-535.

Prelec, D. “A Bayesian truth serum for subjective data.” *Science*, 2004, 306, 462-466.

Surowiecki, J. *The wisdom of crowds*. Anchor, 2005.