The end of an error?

TIMOTHY GOWERS

In 1998 the *Lancet*, one of Elsevier's most prestigious journals, published a paper by Andrew Wakefield and twelve colleagues that suggested a link between the MMR vaccine and autism. Further studies were quickly carried out, which failed to confirm such a link. In 2004, ten of the twelve co-authors publicly dissociated themselves from the claims in the paper, but it was not until 2010 that the paper was formally retracted by the *Lancet*, soon after which Wakefield was struck off the UK medical register.

A few years after Wakefield's article was published, the Russian mathematician Grigori Perelman claimed to have proved Thurston's geometrization conjecture, a result that gives a complete description of mathematical objects known as 3-manifolds, and in the process proves a conjecture due to Poincaré that was considered so important that the Clay Mathematics Institute had offered a million dollars for a solution. Perelman did not submit a paper to a journal; instead, in 2002 and 2003 he simply posted three preprints to the arXiv, a preprint server used by many theoretical physicists, mathematicians and computer scientists. It was difficult to understand what he had written, but such was his reputation, and such was the importance of his work if it was to be proved right, that a small team of experts worked heroically to come to grips with it, correcting minor errors, filling in parts of the argument where Perelman had been somewhat sketchy, and tidying up the presentation until it finally became possible to say with complete confidence that a solution had been found. For this work Perelman was offered a Fields Medal and the million dollars, both of which he declined.

A couple of months ago, Norbert Blum, a theoretical computer scientist from Bonn, posted to the arXiv a preprint claiming to have answered another of the Clay Mathematics Institute's million-dollar questions. Like Perelman, Blum was an established and respected researcher. The preprint was well written, and Blum made clear that he was aware of many of the known pitfalls that await anybody who tries to solve the problem, giving careful explanations of how he had avoided them. So the preprint could not simply be dismissed as the work of a crank. After a few days, however, by which time several people had pored over the paper, a serious problem came to light: one of the key statements on which Blum's argument depended directly contradicted a known (but not at all obvious) result. Soon after that, a clear understanding was reached of exactly where he had gone wrong, and a week or two later he retracted his claim.

These three stories are worth bearing in mind when people talk about how heavily we rely on the peer review system. It is not easy to have a paper published in the *Lancet*, so Wakefield's paper presumably underwent a stringent process of peer review. As a result, it received a very strong endorsement from the scientific community. This gave a huge impetus to anti-vaccination campaigners and may well have led to hundreds of preventable deaths. By contrast, the two mathematics preprints were not peer reviewed, but that did not stop the correctness or otherwise of their claims being satisfactorily established.

An obvious objection to that last sentence is that the mathematics preprints *were* in fact peer-reviewed. They may not have been sent to referees by the editor of a journal, but they certainly were carefully scrutinized by peers of the authors. So to avoid any confusion, let me use the phrase "formal peer review" for the kind that is organized by a journal and "informal peer review" for the less official scrutiny that is carried out whenever an academic reads an article

and comes to some sort of judgement on it. My aim here is to question whether we need formal peer review. It goes without saying that peer review in some form is essential, but it is much less obvious that it needs to be organized in the way it usually is today, or even that it needs to be organized at all.

What would the world be like without formal peer review? One can get some idea by looking at what the world is already like for many mathematicians. These days, the arXiv is how we disseminate our work, and the arXiv is how we establish priority. A typical pattern is to post a preprint to the arXiv, wait for feedback from other mathematicians who might be interested, post a revised version of the preprint, and send the revised version to a journal. The time between submitting a paper to a journal and its appearing is often a year or two, so by the time it appears in print, it has already been thoroughly assimilated. Furthermore, looking a paper up on the arXiv is much simpler than grappling with most journal websites, so even after publication it is often the arXiv preprint that is read and not the journal's formatted version. Thus, in mathematics at least, journals have become almost irrelevant: their main purpose is to provide a stamp of approval, and even then one that gives only an imprecise and unreliable indication of how good a paper actually is.

Of course, different disciplines have different needs and very different publishing cultures. In many subjects, referees typically demand much more substantial changes than they do in mathematics, journals have less permissive policies regarding preprints, and authors are not in the habit of making their work available online. Journals in the biomedical sciences, for example, often do not allow authors to post versions of their articles that have been revised in response to comments from referees. These sometimes differ in important ways from the versions originally submitted: for example, it is not uncommon for referees to go as far as to require authors to carry out further experiments. It also seems that many scientists do not regard posting a preprint as any kind of establishment of priority, and they worry about being scooped if they do so. So, I do not want to propose some unified system that would be used by all academics – a mistake made with depressing frequency by university policy-makers – but I do think that we should each question whether the current system, whatever it might be in our particular discipline, is optimal for that discipline. I also think that in at least some disciplines, my own being an example, the answer is no.

Defences of formal peer review tend to focus on three functions it serves. The first is that it is supposed to ensure reliability: if you read something in the peer-reviewed literature, you can have some confidence that it is correct. This confidence may fall short of certainty, but at least you know that experts have looked at the paper and not found it obviously flawed.

The second is a bit like the function of film reviews. We do not want to endure a large number of bad films in order to catch the occasional good one, so we leave that to film critics, who save us time by identifying the good ones for us. Similarly, a vast amount of academic literature is being produced all the time, most of it not deserving of our attention, and the peer-review system saves us time by selecting the most important articles. It also enables us to make quick judgements about the work of other academics: instead of actually reading the work, we can simply look at where it has been published.

The third function is providing feedback. If you submit a serious paper to a serious journal, then whether or not it is accepted, it has at least been read, and if you are lucky you receive valuable advice about how to improve it.

Let us consider these functions in turn. For each one, we should ask whether formal peer review is the best way of performing that function, and, if so, whether the function is valuable enough to justify the effort and expense of the formal peer review system.

Why does it matter whether academic literature is reliable? One reason is that academics like to build on the work of other academics. In order to convince other academics of the validity of their conclusions (by the standards appropriate to their discipline), they want to be able to draw on an accepted body of knowledge rather than having to justify everything from first principles. Another reason is that non-experts often need to make important decisions based on the conclusions of academics, so they need to know which conclusions have been properly established.

In mathematics, the first reason is particularly important: if I use somebody else's result to prove a theorem, and their result has not been properly proved, then my theorem is not properly proved either. And yet it is not formal peer review that gives me the necessary confidence when I use somebody else's result. Many papers are long and technical, and it is too much to expect a referee to check all the details. As a result, there are many incorrect arguments in the published literature. However, this rarely leads to problems, because the results that are truly important – the ones that are used by many other mathematicians – receive a great deal of informal peer review. Ideally, one's reason for being confident in a result one uses is that one has actually gone to the trouble of understanding how it is proved (which has many benefits besides merely checking correctness). Sometimes that is impractical, but even then what gives one confidence is not the fact that the result has been published, but rather the fact that it has become generally accepted by a community of experts, some of whom have read it carefully and perhaps reworked the argument. And if that is not the case, then one will be anxious about using the result even if it has been published.

In some disciplines, the formal peer-review system appears to have failed on a huge scale. This is particularly true of articles about scientific experiments where the conclusions are statistical in nature. It may look convincing if an experiment yields a result that would have had only a 1 per cent probability of happening purely by chance, but 1 per cent chances happen 1 per cent of the time, so if you do a hundred experiments that look for different effects, you will on average expect to demonstrate one of those effects at the 1 per cent level of statistical significance, even if it does not in fact exist. This process can be made more efficient if one takes several measurements per experiment, since then a single set of data has many more opportunities to give rise to a bogus connection. The current publication system exacerbates this problem, since scientists report only on their positive results. It would be very helpful if they also revealed all their negative results, since that would make it much easier for others to do proper statistical analyses, but there are no rewards for doing so. The result is that in some disciplines, the social sciences being particularly notorious, it has been discovered that a large percentage of the experiments described in the literature, even in the best journals, do not yield the claimed results when repeated.

Nor does formal peer review seem to manage very well to stop wrong ideas from spreading outside academia. Climate change deniers are not put off by their lack of representation in respectable academic journals. Drugs policy bears little relation to the harm that drugs actually cause. An economics paper that supported austerity-based policies influenced several governments before it was discovered to contain some very basic mistakes that invalidated it. Fortunately, this can work both ways: the fact that the articles in the journal *Homeopathy* (formerly the *British Homoeopathic Journal*) have been through a process of formal peer review and are published by a major publisher does not lead to their being taken seriously by non-believers.

Do we need formal peer review in order to help us discover what is worth reading? One might think that if everybody simply put their writings on preprint servers, then the interesting articles would be very hard to find among all the junk. Again, the way things already are in mathematics suggests that this would not be as serious a problem as it at first looks. Far more papers are posted to the arXiv than I have time to check through, even if I restrict myself to a few areas of particular interest. But I use a recently created website called arxivist.com, which puts each day's preprints in what it judges to be their order of interest to me. To get started with the site, I had to answer a few basic questions about my interests, and then I had the chance to tell the system whether I approved of its judgement on any given paper. Based on my feedback, it refined its criteria, and now it does a remarkable job: occasionally I check, and I almost never find a preprint that I want to read that has not made it into the top five or six for the day. That is not to say that I am always interested in the top ones, but it is the work of a moment to glance at their titles and look at the abstracts of the ones that catch my attention. This is a far more efficient process than it would be to keep track of the latest volumes of mathematics journals. The danger with such a system is that I will be confined to my little bubble and will miss developments in areas not my own, but those will usually interest me only if they are particularly noteworthy, and there are other ways of hearing about important breakthroughs, such as blogs and certain corners of social media that are inhabited by mathematicians.

That still leaves the other time-saving aspect: being able to make quick judgements of publication lists by looking at the names of journals. There is no doubt that this practice does indeed save time, but it is highly questionable whether it should, since it encourages the measurement culture that has infected academia, with all its well-known adverse consequences. It is also questionable whether, even if we do want some means of making crude judgements quickly, the formal peer-review mechanism is the best way of providing it.

The third function, providing feedback, is much more obviously valuable, especially in some subjects. Remarkably, we have arrived at a system where academics feel a moral obligation to perform the thankless task of reviewing the work of other academics, anonymously and unpaid. This undoubtedly makes the literature better than it would otherwise have been, and ensures at least one reader for each paper.

It is not hard to think of other systems that would provide feedback, but it is less clear how they could become widely adopted. For example, one common proposal is to add (suitably moderated) comment pages to preprint servers. This would allow readers of articles to correct mistakes, make relevant points that are missing from the articles, and so on. Authors would be allowed to reply to these comments, and also to update their preprints in response to them. However, attempts to introduce systems like this have not, so far, been very successful, because

most articles receive no comments. This may be partly because only a small minority of preprints are actually worth commenting on, but another important reason is that there is no moral pressure to do so. Throwing away the current system risks throwing away all the social capital associated with it and leaving us impoverished as a result.

That is a strong argument against an abrupt change to a new system, but it is not an argument against a gradual one. And it is not unrealistic to hope for gradual change. For example, the habits of mathematicians that I described earlier are very different from those of twenty years ago, but they arose gradually. There was no point at which the rate of posting of mathematics preprints to the arXiv suddenly jumped: rather, it grew steadily from about the turn of the century until it became standard (though not yet universal) practice. If the only problem with a new system is that there is no moral pressure on academics to participate in it, there is the option of attempting to apply such pressure and waiting until it becomes widely felt. Perhaps academics who supported new systems would start agreeing to participate in them while cutting down their activities in the old system.

Why does any of this matter? Defenders of formal peer review usually admit that it is flawed, but go on to say, as though it were obvious, that any other system would be worse. But it is not obvious at all. If academics put their writings directly online and systems were developed for commenting on them, one immediate advantage would be a huge amount of money saved. Another would be that we would actually get to find out what other people thought about a paper, rather than merely knowing that somebody had judged it to be above a certain not very precise threshold (or not knowing anything at all if it had been rejected). We would be pooling our efforts in useful ways: for instance, if a paper had an error that could be corrected, this would not have to be rediscovered by every single reader.

An alternative system would almost certainly not be perfect, but to insist on perfection, given the imperfections of the current system, is nothing but status quo bias. To guard against this, imagine that an alternative system were fully established and see whether you can mount a convincing argument for switching to what we have now, where all the valuable commentary would be hidden away and we would have to pay large sums of money to read each other's writings. You would be laughed out of court.

Source: https://www.the-tls.co.uk/articles/public/the-end-of-an-error-peer-review/