Jury Theorems for Peer Review*

Marcus Arvan† Liam Kofi Bright‡ Remco Heesen§

April 23, 2019

Abstract

Peer review is often taken to be the main form of quality control on academic writings. Usually this is carried out by journals. Parts of math and physics appear to have now set up a parallel, crowd-sourced model of peer review, where papers are posted on the arXiv to be publicly discussed. In this paper we argue that crowd-sourced peer review is likely to do better than journal-solicited peer review at sorting papers by quality. Our argument rests on two key claims. First, crowd-sourced peer review will lead to there being on average more reviewers per paper than journal-solicited peer review. Second, due to the wisdom of the crowds, more reviewers will tend to make better judgments than fewer. We make the second claim precise by looking at the Condorcet Jury Theorem as well as two related, novel jury theorems developed specifically to apply to the case of peer review.

*All authors contributed equally. We thank Justin Bruner, Allard Tamminga, and an audience at the University of Groningen for valuable comments and discussion. RH's research was supported by the Netherlands Organisation for Scientific Research (NWO) under grant 016.Veni.195.141.

†University of Tampa. Email: marvan@ut.edu.

‡London School of Economics and Political Science. Email: liamkbright@gmail.com.

§University of Western Australia and University of Groningen. Email: remco.heesen@uwa.edu.au.

1 Introduction

Peer review is supposed to secure an epistemic benefit for science. By ensuring that only work that has been validated by the judgment of multiple experts is allowed into the scientific literature, peer review is commonly thought to function as a quality control that prevents us from wasting our time on bad work. Rather than have to wade through every half-baked flight of fancy that any old person takes the time to write up, a discerning scientist may simply peruse noted peer-reviewed journals, and read only that which passes the filter of peer review. However, in this essay we argue that scientists (a term we will use broadly to include academics in the social sciences and the humanities) would be better served wading through the half-baked flights of fancy. More exactly, we argue that an open, 'crowd-sourced' approach to peer review is more likely to reliably identify high-quality work compared to traditional, journal-solicited peer review.

The standard system of journal-solicited approach to peer-review widely practiced today filters the quality of academic work through a small number of experts—typically, a small group of editors and anywhere from one to three outside referees who read and evaluate (typically) anonymized submissions. The normative assumption that appears to underlie this practice is the belief that a small number of experts (reading anonymized submissions) are *the best mechanism* for distinguishing between high-quality and low-quality work, and hence, for determining which articles should appear in academic journals. Notice that in this system, quality assessment occurs in two stages: first, in pre-publication peer review, where a small number of experts determine which papers should be published in a given journal; and then, second, by a journal's readership post-publication. We take it that the first of these two stages is intended to provide a proxy for the second, i.e., the more long-term assessment by the field. In this paper, we argue that a 'crowd-sourced' approach to peer review that bypasses the first stage—immediately opening up papers for evaluations by large numbers of readers, inclusive of both

experts and readers that a contemporary journal editor might not regard as experts—is likely to *more reliably* evaluate paper quality than the traditional model of peer review. In brief, we will appeal to the Condorcet jury theorem (Condorcet 1785) and some closely related, novel mathematical results to argue that a large number of evaluators is more likely to converge on an accurate quality assessment of a paper than a smaller number of evaluators.

We will go into more detail regarding the system we advocate for in the following sections. For now, some remarks on how this relates to the previous literature. There has been a revival of interest in the epistemic benefits to be secured by large numbers of diverse agents (List and Goodin 2001, Hartmann and Sprenger 2012), including in the social epistemology of science (Heesen et al. forthcoming, O'Connor and Bruner 2019, Singer 2019). Our intent is to bring this literature to bear on a concrete and applied problem in the social epistemology of science, namely the implementation of peer review. Further, given the replication crisis, there has recently been interest in systematic failures of the peer review system (Romero 2016, Heesen 2018). Our paper could also be seen as offering a thoroughgoing solution to the problems there identified. Like those who argue we should do away with peer review in the context of project funding (Avin forthcoming), we think we should do away with the idea that a small number of experts can reliably predict in advance which ideas will be worth pursuing or reading. We are not the first to suggest opening up peer review (e.g., Nosek and Bar-Anan 2012), but we offer a novel argument for its epistemic benefits. If our argument is sound, *nothing* should get a potentially deceptive stamp of authority through pre-publication peer review. Instead, to gauge the quality of scientific work, we should rely on the long-run and aggregated views of many scientists responding to the work of their peers through their diverse inquiries.

2 Assumptions of Peer Review

Our purpose is ultimately to compare and contrast different ways of arranging peer review. In the present paper the different arrangements being considered are, on the one hand, the present system of journal-solicited pre-publication peer review, and on the other hand an alternative system of open, crowd-sourced, post-publication peer review. However, we expressly do not engage in an all-things-considered comparison of these two systems. Rather, we focus on one particular goal that peer review is intended to achieve (one that we think many would take to be the most important goal of peer review), namely the selection of high-quality papers.

In this section we argue that *if* the present system of peer review really helps us pick out high-quality papers then scientific quality and the peer reviewers who assess it must satisfy certain features. The structure of our argument here is loosely analogous to a transcendental deduction: we argue that without satisfying these features, the idea that any form of peer review could successfully select for quality does not make sense. In subsequent sections we argue that if quality and reviewers really have these features, then a system quite distinct from journal-solicited peer review would do better at differentiating high-quality and low-quality scientific work. Hence, our overall argument is that if the necessary conditions are in place for journal-solicited peer review to select for quality, we ought not to make use of journal-solicited peer review.

Our focus in this section is on three assumptions which we take it are necessary features of peer review if it is to select for high-quality papers. In particular, we take anyone who defends the present system of peer review on the grounds that it is a (however imperfect) way to filter the scientific literature on the basis of quality to be committed to these assumptions. The three assumptions are *competency*, *intersubjectivity*, and *atomism*. We give each of these some informal defense.

The first of these is the competency assumption. We assume that scien-

tists are at least better than chance at picking out papers of high quality, or maybe at ranking papers according to their quality, depending on what one takes quality to be or the tasks of reviewers to be. What is significant for our purposes is just that quality is the sort of thing which a peer reviewer is capable of discerning and responding to. If this were false, then the current system could do no better than a system of random publication . So quality is the sort of thing which scientists can and do discern, and respond appropriately to given their reviewing task.

The second assumption is that there exists broad (if rough) intersubjective agreement about what constitutes quality. Here the idea is that for any given scientific paper, there is a unique notion of quality: one way of being the best version of that paper which any reader or reviewer (normatively) should pick up on. Combined with the competency assumption, this would have the consequence that there is a unique notion of quality for any given paper that peer reviewers in fact can pick up on.

Note what we are not saying here: we are not saying that there is only one type of quality, and that peer reviewers should always be picking up on that. Rather, we are assuming that once you fix facts about a paper's topic and the type of impact it is intended to have *then* it turns out that there is a unique best way of fulfilling the paper's purpose.

To see why this assumption underlies present peer review practices, consider what would be the case if it were false. If quality is not intersubjective then it is unclear what peer review is doing. Why should the fact that these reviewers like it give me reason to think that it will be high quality in the sense pertinent to me? It would always be possible that while the paper does very well on some notion of quality peer reviewers are responding to, this is systematically failing to pick out what readers actually care about (or ought to care about) when deciding how to allocate their time among scientific journals. The fact that we assume that peer reviewers can assess quality and in that way make useful judgments about what is worth spending time upon

belies a presupposition that for any given paper the relevant experts know what it would take to be a more or less worthy version of that paper, and that we can reasonably expect some degree of agreement on this point. We allow, and in fact it will be essential to our argument, that this agreement may be partial and may be accompanied by substantive and persistent disagreement on particular points.

Finally, take the notion of atomism. Say that paper quality is atomistic if one can discern a paper's quality by means of evaluating just that paper. Here we might think, in particular, of quality as representing something like the marginal contribution of a new paper: one may take into account one's knowledge of what already exists in the literature, since peer reviewers surely have some such information. However, to assess a new paper they need only look at that paper, as opposed to some bundle of papers which could potentially be released at once.

That this is presupposed seems to follow from how we arrange things. Outside of special editions of journals, editors make decisions on a case-by-case basis. Further, peer reviewers usually are not given any information about what else is in the pipeline, or what sort of other papers the item they are assessing could be paired with. Rather, they are asked to issue judgment on a given manuscript by itself. If paper quality were not atomistic this procedure would be ill-advised, since editors and peer reviewers may simply be ignoring pertinent information for assessing the paper's quality.

Hence we assume that if peer review is to make sense as a system of quality control the following features must be in place. Paper quality is such that for any given paper there is a unique best way it could fulfil its own potential; the degree to which it does this can be assessed by people reviewing just this paper (in light of their background knowledge of this literature) rather than requiring any more holistic judgement; and the scientific community contains people who are indeed competent to pick it out with just these features. We shall presently construct a model with these features, and argue that journal-

solicited peer review would not represent the best way of getting at quality under these assumptions.

While our focus in this paper is exclusively on the role of peer review in quality control, we feel it is useful at this stage to briefly discuss *fairness* considerations. Proponents of journal-solicited peer review often suggest that our current practices are the fairest method available for selecting papers for publication. A central contention here is that a particular feature of the existing peer review process—anonymization—is vital for protecting against reviewer bias. We take no stance on anonymization. Our arguments instead support the following conditional claims: if anonymization is important for fairness in peer review, then an anonymized crowd-sourced peer review process would be superior to current processes; and if anonymization is unimportant, then a non-anonymized crowd-sourced peer review process would be superior. Either way, our point stands: an open, crowd-sourced method of peer review is likely to more accurately judge paper quality than journal-solicited peer review.

3 The Basic Condorcet Jury Theorem

In the previous section, we argued that if peer review is to reliably select high quality papers for publication (as proponents of the practice allege it does), then three assumptions must hold. First, referees must on average be *competent* to judge paper quality better than chance. Second, paper quality itself must be *intersubjective*, such that there is some real property of paper quality that competent reviewers track. Third, paper quality must be *atomistic*, such that the quality of a paper can be evaluated by an evaluator considering that paper alone, given the evaluator's background knowledge of what already exists in the literature. The question we ask is whether—granting these assumptions—our current system of journal-solicited peer review is the best available system for selecting high-quality papers for publication. In

this section and the next two, we will provide formal arguments that it is not, and that a crowd-sourced peer review model—of a sort already utilized in academic mathematics and physics—is likely to be a superior method for assessing paper quality.

Our argument in this section is an application of the Condorcet Jury Theorem. One notable feature of journal-solicited peer review is that paper quality is judged by a comparatively small number of evaluators. First, papers are often read and 'desk-rejected' by a single editor. Second, when papers are sent out for review, they are typically reviewed by anywhere from one to three referees. Contrast this system to the crowd-sourced peer review system in math and physics. In these disciplines, it is standard for unpublished papers to be posted on individuals' professional websites and on central repositories, such as the arXiv. It is also a disciplinary norm for members of the academic profession to read and publicly evaluate new submissions on well-trafficked weblogs and repository message-boards. Consequently, the most central part of the peer review process—evaluating paper quality—is more widely distributed. If it is generally expected in an academic profession—as it is in math and physics—that unpublished preprints should be read and discussed publicly before publication, then chances are high that under crowd-sourced peer review the average paper will be reviewed by a larger number of reviewers than in a journal-solicited system.

It is worth noting that this does not necessarily require the overall time spent reviewing papers to increase. Suppose instead that the current disciplinary norm to volunteer one's time to review papers for journals is shifted over to the new system of crowd-sourced peer review, in such a way that each member of the academic community volunteers exactly the same amount of time and reviews the same number of papers. Under journal-solicited peer review (at least as currently implemented), journals base their decision about the quality of a paper only on the reviews solicited by that journal. For those papers that have already been previously rejected from other journals, all pre-

vious reviews are ignored by the journal. In contrast, under crowd-sourced peer review all reviews are public. Thus, if the total number of reviews stays the same, the average number of reviews per paper under crowd-sourced peer review will be higher than the average number of reviews per paper that any given journal has access to.

We will assume throughout the rest of this paper that moving to crowd-sourced peer review increases the average number of reviews per paper. For a reader who thinks, despite our claims above, that crowd-sourced peer review will lower the number of reviews per paper, the rest of the present paper may still hold something of value: the arguments presented below can then be read as arguments in favor of journal-solicited peer review over crowd-sourced peer review. Further, they could be read as a normative argument in favor of increasing the average number of evaluators of any given paper.

Now consider the Condorcet Jury Theorem. The Condorcet theorem shows that, subject to certain assumptions, the judgments of a jury—a group charged with voting on the truth of a proposition (where a majority vote wins)—will have a greater probability of accuracy the greater the number of people included in the jury. The theorem is based on three assumptions. The first assumption is that, with respect to the proposition the jury is judging, there is a *correct answer*: the proposition the jury is judging is either true, or it is false. The second assumption is that every member of the jury has some *independent probability*, p, of voting for the correct truth-value of the proposition under consideration. Finally, the Condorcet theorem is based on the assumption that the *average* probability that any individual in the jury votes correctly is greater than .5. The theorem then says that adding more voters to the pool *makes it progressively less likely* that the majority vote will point to the wrong conclusion.

Here is a brief, intuitive illustration of the theorem. Suppose the average probability that a person selected to a jury will vote for the right answer is $p = .51$. If only 100 people are selected to serve on the jury (and the three

assumptions above hold), then the most likely result is that 51 jury members will vote for the correct answer and 49 for the wrong answer. However, as with all matters involving chance, it is *possible* for the result to deviate from this—as each jury member's likelihood of voting for the right answer is only slightly better than chance ($p = .51$). Because each jury member's probability of voting correctly is just that—a probability—there is a .49 probability that each jury member will vote for the *incorrect* answer. If, for example, just *one* additional jury member votes wrongly, then instead of a 51-49 majority vote for the correct verdict, the jury will vote for a 50-50 tie between the right verdict and wrong verdict. And, of course, if *two* additional jury members vote wrongly, then the jury will vote 51-49 for the *wrong* verdict. Because it only takes a couple of additional wrong votes to go from the single most likely outcome (51 correct votes) to an outcome in which the majority votes wrongly, it is not unlikely for a jury of 100 to go wrong (this happens with a probability of approximately .38). Now, however, consider a jury pool of $100,000$. If each jury member's probability of arriving at the correct verdict is the same $p = .51$, the most likely result is that $51,000$ jury members will vote for the right verdict and only $49,000$ for the wrong verdict. Consequently, in contrast to the first jury, where only *one or two* jury members' mistakes are needed to shift the entire jury's vote to the incorrect verdict, in this case a *thousand* jury members would have to make a mistake to shift the jury's verdict (from the single most likely outcome) to the wrong result. But this is highly unlikely. In fact, the probability of a wrong verdict with a jury of $100,000$ is of the order 10^{-10}, or one in ten billion.

The point generalizes: the larger the jury, the *more accurate* the jury's majority vote is likely to be. In fact, the Condorcet theorem also shows that in the limit case (an infinite-sized jury), the majority will vote for the correct answer with probability 1 (i.e., 100 % of the time). The relevant thing for our purposes, however, is the *comparative* claim: the proposition that the more jury members there are, the more likely it is that a majority of them

will vote for the correct answer. This is important because even if the typical article in a crowd-sourced peer review model is only read and evaluated by a relatively small number of readers (say, four or five members of the academic profession), it is still likely that it will be read and evaluated by *at least as many* independent evaluators as any given paper under journal-solicited peer review.

To see how the Condorcet theorem plausibly supports crowd-sourced peer review, compare the three assumptions of the theorem to the three assumptions discussed in §2. The Condorcet theorem's first assumption is that, with respect to the proposition the jury is judging, there is a *correct answer*. In the case of peer review this proposition would be something like 'This paper is of high quality'. Notice the close connection with the first assumption we argued peer review must satisfy in order to reliably select papers on the basis of quality: *intersubjectivity*, the assumption that there is an intersubjective quality standard for a paper on a particular topic. Notice also that for the moment we are assuming that peer reviewers give (only) a binary judgment of quality: thumbs up or thumbs down. One of the motivations of the models in §4 and §5 is to consider more informative, graded reviewer judgments.

Now consider the Condorcet theorem's second assumption: that every jury member has an *independent probability* (p) of voting for the correct result. Compare this to the second assumption we argued peer review must satisfy: *atomism*, the assumption that the quality of a paper can be evaluated by considering that paper alone. While there is a superficial similarity, these assumptions are quite different: atomism says papers can be evaluated independently of other papers, whereas the assumption we need for the Condorcet theorem is that each reviewer's evaluation is independent of other reviewers. Under crowd-sourced peer review, we can imagine reviewers' judgments becoming correlated (i.e., not independent) due to reviewers being able to read other reviews, whereas under journal-solicited peer review, the active hand of an editor may likewise induce correlation of reviewers' judgments.

So whether the independence assumption is satisfied may well depend on the mechanism by which the different peer review systems are implemented. We will say more about steps a crowd-sourced peer review model could take to ensure reviewer independence in §6.1 and §6.3, but for now we emphasize that independence is assumed in the Condorcet Jury Theorem, and hence the real-world applicability of our argument in this section hinges on providing a mechanism to guarantee it.

Finally, consider the Condorcet theorem's third assumption: that on average voters' probability of voting for the correct answer is better than chance. This corresponds to the third assumption we argued that peer review must satisfy: *reviewer competence*, the assumption that reviewers are capable of picking out high-quality papers, or at minimum that they do better than chance at doing so.

So our claim is that for peer review to 'work as advertised'—that is, for peer review to reliably select papers for publication on the basis of quality—two of the three assumptions of the Condorcet Jury Theorem must be satisfied. Moreover, the third assumption (independence) will be satisfied by crowd-sourced peer review if it is carefully implemented (where we have deferred our discussion of what this means to §6). But then a crowd-sourced peer review model is more reliable than our current journal-solicited peer review model. For whereas journal-solicited peer review is based on the judgments of a *small* jury (usually one editor and one to three reviewers), a crowd-sourced peer review model will—provided appropriate disciplinary norms are in place—tend to base paper evaluation on the judgments of an *equal-sized or larger* jury. And by the Condorcet Jury Theorem a larger jury is more likely to arrive at an accurate evaluation of a given paper's merits.

4 A Jury Theorem for Reviewer Scores

In the previous section we argued that the crowd-sourced method of peer review is superior to the present system on the basis of an 'off the shelf' application of the basic Condorcet Jury Theorem. While we find this argument convincing, we recognize that the basic Condorcet model is highly idealized and as such has a number of features that a skeptic may want to object to. In this section we provide a new model, intended to be more tailored to the specifics of peer review, and show that an analogous theorem holds in this model. This shows that the jury theorem is robust against certain changes in its assumptions, thus strengthening the argument from the previous section.

The basic Condorcet model assumes that agents make a binary judgment on a single proposition. In contrast, real peer reviewers (whether crowd-sourced or journal-solicited) usually provide more nuanced judgments. These may come in the form of numerical scores or in the form of qualitative reasons for the reviewer's verdict. This section considers a model of peer review where reviewers only provide a numerical score; we will add qualitative reasons to the model in the next section.

Whereas in the previous section the goal was to evaluate the truth value of the proposition 'This paper is of high quality', now the goal is to estimate the (numerical) quality of a paper. By the intersubjectivity assumption, there is a particular value that can accurately be said to be *the* quality of the paper, which we will represent by a real number q.

Each review consists of a numerical score, which is the reviewer's estimate of the quality of the paper. We write q_i for the quality estimate (score) provided by reviewer i. By the competence assumption, there is some correlation between reviewer scores and the real intersubjective quality q. But as in the previous section, we will assume that there is some random variation in reviewer scores, reflecting individual reviewer biases and idiosyncrasies. Also as in the previous section, we assume that this variation is independent across reviewers, so reviewer scores can be modeled as independent random

draws from a large pool of potential reviewers or reviewer scores (we refer again to §6.3 for more discussion of the independence assumption).

In this setup, we can represent the competence assumption by assuming that, on average, reviewer scores are equal to the intersubjective quality (that is, $\mathbb{E}[q_i] = q$ for all i). And we can represent reviewer biases and idiosyncrasies by assuming that there is some random variation around this average ($\mathrm{Var}[q_i] = \sigma^2 > 0$ for all i).

Given differing quality estimates from reviewers that are each taken to be competent, it seems reasonable for a journal editor or a reader on the arXiv to take the average of these estimates to be her best estimate of the quality of a paper. Averaging in this way has been defended in the literature on combining forecasts (Clemen 1989, Armstrong 2001, especially p. 422) and peer disagreement (Elga 2007, Christensen 2007, Cohen 2013), while (weighted) linear averaging more generally has also been widely defended by formal epistemologists (Lehrer and Wagner 1981, Martini and Sprenger 2017, Pettigrew 2019). So the quantity of interest that will be used to make decisions under either journal-solicited or crowd-sourced peer review is the average reviewer score $\frac{1}{n}\sum_{i=1}^{n} q_i$.

Because individual reviewer scores are expected to be equal to the intersubjective quality, so is the average reviewer score (that is, $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} q_i] = q$). Perhaps more importantly, the random variation in the average reviewer score will decrease as the number of reviewers increases, according to the formula $\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} q_i] = \sigma^2/n$. This means that, the more reviewers there are, the smaller the probability that the average reviewer score will be much different from the intersubjective quality that it tries to estimate.

This gives us a clear analogy to the Condorcet Jury Theorem in this model with numerical reviewer scores: previously, increasing the number of reviewers increased the probability of a correct verdict, whereas here increasing the number of reviewers increases the probability that the quality estimate obtained from their reviews is close to the correct value. To complete the

analogy, note that the random variation will reduce to zero in the limit as the number of reviewers becomes infinite, meaning that the average reviewer score will be equal to the intersubjective quality with probability one.

Once again, granted the assumption that crowd-sourced peer review will have on average more reviewers per paper than journal-solicited peer review, this yields an argument in favor of crowd-sourced peer review. Namely, crowd-sourced peer review fulfills the objective of selecting high-quality papers better than journal-solicited peer review, as it is more likely to yield accurate quality judgments.

5 A Jury Theorem for Reviewer Reasons

In this section we expand on the model of the previous section by including reviewers' reasons for giving a particular (numerical) quality judgment. We represent these reasons by thinking of papers as having a number of features and peer reviewers as having opinions on which combinations of features make for a high-quality paper. More specifically, we assume there are m features that peer reviewers evaluate for a paper on a certain topic (recall that we allow quality standards to be paper-specific). A paper is represented in the model by its feature coordinates $x_1, x_2, \ldots, x_m$, which provide a numerical 'score' for how that paper does on each of the features. We imagine that for each feature there is a kind of 'golden mean' (possibly relative to the value of the other features) such that both more and less of that feature would make the paper worse in the eyes of the reviewer.

For example, say that for a given paper feature 1 concerns the paper's discussion of the external validity of its results, so x_1 is a number indicating how the paper scores on this feature. A low value of x_1 might indicate that the discussion constrains the external validity too narrowly (compared to what might be justified as represented in the scores on other features); a high value could then be interpreted as saying that the discussion generalizes the study's

results in too widely (i.e., in ways that are not sufficiently supported by the evidence); whereas a medium value indicates a sensible discussion with an eye on applications but avoiding wildly unsupported claims.

What might the set of features look like? Since the model works for an arbitrary number of features, we can remain somewhat agnostic about this, but here are some suggestions. First, the features might be Kuhn's criteria for theory choice: empirical adequacy, simplicity, etc. Second, the features might be some variation on those that peer reviewers are explicitly asked to score papers on by journals: novelty, methodological soundness, etc. Third, the features might be anything and everything peer reviewers use to evaluate papers, at as fine-grained a level as possible (the example in the previous paragraph gives an indication of how fine-grained this might be). We prefer this third option, as it gives the most detailed and realistic analysis that is possible within this modeling framework.

According to our intersubjectivity assumption, there is such a thing as the intersubjectively agreed quality of a given paper. We conceive of both the intersubjective quality of a paper and any given reviewer's opinion of its quality as a function of that paper's feature coordinates $x_1, \ldots, x_m$. Quality can then be characterized as something that looks like an epistemic landscape (in the sense of Weisberg and Muldoon 2009, Alexander et al. 2015, Thoma 2015): each m-dimensional point $(x_1, \ldots, x_m)$ represents a possible paper and the height of the landscape at that point is the quality of a paper with such characteristics. In particular, we define the function $f : \mathbb{R}^m \to [0, \infty)$ to describe the epistemic landscape corresponding to intersubjective quality. That is, $f(x)$ is the intersubjective quality of a paper with characteristics $x = (x_1, \ldots, x_m)$.

In accordance with our competence assumption, peer reviewers (whether crowd-sourced or journal-solicited) are in a good position to estimate the quality of a paper. However, they are not perfect at this. First, they may be biased in the sense that the combinations of features they perceive to

indicate high quality are slightly different from the combinations that really constitute intersubjective quality. And second, there may be measurement error in determining the feature coordinates, i.e., peer reviewers may make mistakes in estimating where a paper falls on the scale for some or all features. We roll these two types of errors into a single bias b_i for a given reviewer i. The bias b_i is an m-dimensional point representing the total distortion in reviewer i's estimation of quality due to these two types of error, such that the reviewer's quality estimate for a paper with characteristics x will be $f(x - b_i)$.

We use μ to denote the center of mass of the epistemic landscape of intersubjective quality, and we assume that this quantity exists.[1] As a consequence, for any reviewer i, the epistemic landscape characterizing how that reviewer estimates quality also has a center of mass, and it is located at $\mu + b_i$.

As in the previous section, we assume that a journal editor or a reader on the arXiv takes the average of these estimates to be her best estimate of the quality of a paper. For a paper with characteristics x reviewed by n reviewers we will denote this estimate $f_n(x)$. Since the quality estimates from the reviewers are $f(x - b_1), f(x - b_2), \ldots, f(x - b_n)$, the editor's or reader's estimate of the paper's quality will be $f_n(x) = \frac{1}{n} \sum_{i=1}^{n} f(x - b_i)$.

Also as before, we assume that crowd-sourced peer review will lead (on average) to more reviewers per paper than journal-solicited peer review. The question we are then interested in is whether a greater number of reviewers will improve the quality estimate, i.e., bring it closer to the intersubjective quality. For a given paper x, this translates in the model to the question whether $f_n(x)$ gets closer to $f(x)$ as n increases. Depending on the shape

[1] More formally, we assume that $\int_{\mathbb{R}^m} x_j f(x)\, dx$ is finite for each feature j and then we define μ coordinate-wise by $\mu_j = \int_{\mathbb{R}^m} x_j f(x)\, dx / \int_{\mathbb{R}^m} f(x)\, dx$. Given our 'golden mean' approach to paper quality this assumption is fairly innocent. In particular, if f has a finite maximum (as it does under any reasonable formalization of the 'golden mean' approach) and the features are measured on finite scales then the assumption definitely holds. If the features are measured on infinite scales the assumption may still hold but it will depend on how quickly quality drops off as you move away from the maximum.

of the landscape and reviewers' biases, this may be true for some values of x and false for others. What we would like to know more specifically, then, is whether for an arbitrary paper the quality estimate gets closer to the intersubjective quality with more reviewers, i.e., whether the function f_n as a whole becomes more similar to the function f as n increases.

How do we characterize the similarity of two functions? Here we will take the following approach: compare the center of mass of each function. The center of mass measures the central tendency of a function, giving some indication of where in the landscape the highest peaks of quality occur. This is a fairly crude measure of similarity, as two functions may have the same center of mass but still be quite dissimilar in other respects, but it has the advantage of giving us a single number (or m-dimensional point, to be more precise) for each function. This measure works well when the landscapes under consideration are single-peaked and mostly smooth, as in such cases two landscapes with similar centers of mass will usually agree on relative quality judgments (i.e., on which of two points in the landscape makes for a better paper). The center of mass of f_n is $\mu + \frac{1}{n}\sum_{i=1}^{n} b_i$, which we want to compare to μ, the center of mass of f.

Now we just need to worry about how the reviewer biases b_i are distributed. We assume that we can treat these as random variables, i.e., that reviewer selection can be viewed as selection from a larger population of potential reviewers governed by some probability distribution. This need only be true in a subjective sense: the bias of a given reviewer is random insofar as you don't know in advance which reviewer and hence which bias will be selected. We assume that reviewers are chosen in such a way that expected bias is zero (speaking somewhat loosely, this says that bias is equally likely to be in any direction) and expected variation in this bias is finite.[2]

[2]More formally, we assume (for all i) that $\mathbb{E}[b_i] = 0$ and $\mathrm{Cov}[b_i] = \Sigma$. Here, 0 is the m-dimensional vector of zeroes, and Σ is the covariance matrix of a particular reviewer's bias. That is, Σ gives the covariances between the reviewer's bias in each of the m features, not the covariances between different reviewers' biases. In virtue of being the covariance ma-

It follows that in expectation the center of mass of estimated quality is equal to the center of mass of intersubjective quality ($\mathbb{E}[\mu + \frac{1}{n}\sum_{i=1}^{n} b_i] = \mu$), that is, on average there will be no bias at all regardless of the number of reviewers. But assuming the biases of different reviewers are independent[3] we also get that the probabilistic variation in the center of mass of estimated quality decreases with the number of reviewers ($\mathrm{Cov}[\mu + \frac{1}{n}\sum_{i=1}^{n} b_i] = \Sigma/n$). This means that the center of mass of estimated quality is more likely to be far away from the center of mass of intersubjective quality if there are fewer reviewers, and more likely to be close if there are more. Moreover, since the variation reduces to zero in the limit, $\mu + \frac{1}{n}\sum_{i=1}^{n} b_i$ probabilistically converges to μ.

These results provide a close parallel in our model to the basic Condorcet Jury Theorem. We have a result saying that things get 'better' (estimated quality is likely to be closer to intersubjective quality) as the number of reviewers increases, and we have a guarantee that things are 'good' (estimated quality as close to intersubjective quality as it can get by this fairly crude measure) in the infinite limit. We conclude from this that crowd-sourced peer review, insofar as it tends to involve a greater number of reviewers, outperforms journal-solicited peer review even when we grant the basic assumptions (competency, intersubjectivity, and atomism) that are required for journal-solicited peer review to make sense.

6 Replies to Potential Objections

We anticipate several objections to our argument, each focusing on a different background assumption.

---

trix, Σ is a symmetric and positive semi-definite $m \times m$ matrix. We make no assumptions on Σ other than that it is not the zero matrix.

[3]Note that we have allowed that reviewers' biases may be correlated in different features: the off-diagonal elements of Σ may be nonzero. But the independence assumption here rules out correlations between the biases of distinct reviewers. For discussion of this assumption we refer again to §6.3.

6.1 Manipulation of Reviewer Scores?

A number of our assumptions (reviewer competence and independence of reviewer judgments in particular) are not even remotely plausible for crowd-sourced peer review if it is overwhelmed by, say, internet trolls with a political agenda, or other forms of organized manipulation. If people are basing their judgment of a paper on reasons having nothing to do with the quality of a paper (e.g., political reasons) then the reviewer competence assumption will not be satisfied. And if groups of people are mobilized to leave reviews of a paper without putting much or any of their own thought into it then the independence assumption will not be satisfied. Note that the latter will be a problem for our independence assumption regardless of whether such a 'mass reviewing campaign' is ultimately motivated by scientific (e.g., a large research program ganging up on a smaller one), political, or other reasons. Given the extent to which various social media have recently been overwhelmed by these types of phenomena, an important worry about crowd-sourced peer review is whether mechanisms can be put in place to prevent the same from happening there.

It is tempting to try to address such issues by putting tight restrictions on who is allowed to review. There are a number of ways of doing this. We might use formal requirements such as possession of a doctorate or being employed by a university, or social requirements such as having to be endorsed by existing reviewers or having to have your reviews be rated sufficiently helpful by other reviewers. And we might apply such requirements at a system-wide level (i.e., to decide whether a given person is allowed to review anything at all) or at a subfield-specific level, e.g., requiring a doctorate in a specific subfield or endorsement from other reviewers in that subfield to be allowed to review papers in that subfield.

However, to our mind such restrictions go against the spirit of the proposal of crowd-sourced peer review and will severely limit some of its key advantages. Restricting who is allowed to review will inevitably lower the

average number of reviews per paper, thus reducing the benefits from large numbers we have discussed at length in the preceding sections. Moreover, the particular restrictions suggested above will reinforce existing disciplinary boundaries and subfield-level groupthink (where it exists), whereas one of the key strengths of crowd-sourced peer review as we imagine it is that it will be easier for disparate fields to cross-pollinate, benefit from each other's insights, and correct each other's biases.

For these reasons we favor a system in which anyone is allowed to review anything, regardless of whether they are a recognized expert on the particular topic, and regardless even of whether they are an academic at all. But this does not mean giving free rein to trolls, mobs, and other forms of manipulation. We think there are various measures that can be put in place to guard against this.

Here is a relatively simple one to implement. For each subfield, curate a set of expert reviewers along the lines suggested in the previous paragraphs, e.g., have reviewers endorse each other's expertise in the given subfield. Here we imagine subfields to be relatively small, such that there will be more than twenty but less than a hundred endorsed reviewers per subfield. Now, for each paper, report both the overall average reviewer score and the average reviewer score when taking into account only reviewers endorsed for that particular subfield.

This system, familiar from the film review website Rotten Tomatoes, has a number of advantages. Conservative-minded scientists who prefer something close to traditional journal-solicited peer review can take into account only the endorsed reviewer average, whereas those who put their faith in the wisdom of the crowds can focus on the overall average. But perhaps more importantly, one can look at both. When they are relatively similar, either there were little or no non-expert reviewers, or the non-expert reviewers tended to agree with the endorsed reviewers. But it gets interesting when there is significant divergence between the endorsed reviewer average and the

overall average. This could be evidence of a mob coming in to manipulate the score, but it could also be evidence of groupthink within the subfield being exposed by the independent insights of outside experts. In any case, the divergent scores will be a signal that something is up (Nosek and Bar-Anan 2012, p. 238). Individual readers will be alerted that at least one of the scores is misleading, and that blind reliance on the averages is not advisable for this particular paper.

More generally, while the specific argument of this paper is that using the overall average reviewer score from crowd-sourced peer review will give better quality judgments than can be obtained using journal-solicited peer review, it is emphatically not part of our proposal that overall average scores should be the only thing available to academics in deciding what to read. Our view is rather that a lot of additional information should be made available to potential readers, so they can freely choose which metrics they think are more informative. This includes, but is not limited to, the content and score of each individual review, the total number of reviews and the number of reviews by endorsed reviewers, the ranking of the paper relative to other papers in its subfield, as well as potentially things like ratings and rankings of the helpfulness of individual reviewers. Combining some of these metrics will provide additional insight relevant to the problem of manipulation. For example, one would normally expect better papers to receive not just better reviews but also more reviews as more readers are attracted by the high score. Thus, a paper with an unusually high number of reviews but a low average score should raise suspicions. Same thing for a paper where most of the reviews come from reviewers who have never reviewed anything else. Using all this information, we think academics will be able to make use of crowd-sourced peer review to identify and read high-quality papers with minimal interference from manipulation.

6.2 Greater Average Competence in Journal-Solicited Peer Review?

Our arguments that a crowd-sourced peer review system is more likely to produce accurate group judgments of overall article quality than traditional journal-solicited peer review are based on an assumption that *reviewer competence* is randomly distributed throughout the population of possible reviewers. That is, in the argument of §3 we assume that the (average) probability p that a reviewer in journal-solicited peer review will arrive at an accurate judgment of a paper's merit is the same as the (average) probability of an accurate judgment from a crowd-sourced reviewer. And similarly, in the argument of §5 we assume that the bias of a randomly selected peer reviewer follows the same probability distribution regardless of whether the reviewer is journal-solicited or crowd-sourced.

However, some may doubt this equivalence. First, some might suggest that editors at journals are likely to select substantially more competent reviewers than the average reviewer in the population. Editors at the most selective, highly ranked journals in particular might seek out and commission reviews by the most accomplished figures in the field (in philosophy, eminent figures such as Timothy Williamson, Christine Korsgaard, etc.).[4] These reviewers, due to their exceptional achievements, may perhaps be expected to have a higher probability of accurately judging a given paper's merits or be systematically less biased (as expressed, say, in a lower variance in the bias defined in the model of §5). Second, even if this were not the case, some might argue that insofar as journal-solicited peer review commis-

[4]See `https://philosopherscocoon.typepad.com/blog/2018/12/incentivizing-better-reviewer-behavior.html`, where David Bourget notes that journal editors may aim to select more accomplished, senior scholars as reviewers, the assumption being such figures are more qualified to review a paper than the population of scholars at large. As Bourget puts it, "Bearing in mind that the relevant junior/senior distinction isn't age-based but accomplishment-based, it seems natural to expect that senior reviewers are on average at least a little more skilled and have beneficial experience and perspective when it comes to assessing new work."

sions reviews by specialists in the paper's field (e.g., political philosophers for political philosophy papers, etc.), those specialists are likely to have a higher probability of accuracy or be less biased than a pool of reviewers that includes non-specialists. If as a result of either of these two mechanisms journal-solicited peer review reliably selects reviewers who are more accurate or less biased than crowd-sourced peer review, then our arguments do not go through. In order to successfully defend a Condorcet-style argument in this case, we would need to show that the *accuracy increase* generated by increasing the size of the jury pool (through crowd-sourced peer review) is *greater* than the accuracy increase generated by how journal editors select reviewers. But this we have not shown—thus, the objection goes, our defense of crowd-sourced peer review is unsound.

Our reply to this concern is two-fold. First, the balance of present evidence suggests the empirical claim that journals select better reviewers is not true. Second, there are a number of *prima facie* reasons to believe that journal-solicited reviewers are likely to be *more* biased than the population from which crowd-sourced peer review might draw.

Although it is sometimes suggested in informal discussions that journal-solicited reviewers are likely to be more competent to evaluate papers than readers at large[5], is there any reason to believe this is true? Notably, arguments that it is tend to come from the armchair—from the intuitive thought that journals 'ask the best people' to review papers, and the 'best' people are likely to be the most accurate judges of paper quality. However, two sources of empirical evidence collectively cast doubt on these claims. First, empirical studies that have been done on the quality of journal-solicited reviewers suggest very low interrater reliability (Lee et al. 2013, pp. 5–6; Bornmann 2011, p. 207). Interrater reliability measures the level of agreement between different reviewers judging the same paper. This is relevant here because lots

[5]Again, see `https://philosopherscocoon.typepad.com/blog/2018/12/incentivizing-better-reviewer-behavior.html`.

of disagreement between reviewers entails that individual reviewer accuracy cannot be particularly high. In one study, interrater reliability barely exceeded chance (Kravitz et al. 2010). In terms of the basic Condorcet model, this corresponds to probabilities of voting correctly barely exceeding .5.

Second, there are wide-ranging anecdotal reports that reviewers at highly-selective journals *routinely misjudge* papers that much larger audiences have judged more accurately. Remarkably, Gans and Shepherd (1994) reported how a wide variety of classic (including Nobel Prize winning) economics articles were systemically rejected by top-ranked journals in the field. At least anecdotally, this also happens in academic philosophy—where, for instance, Jason Stanley reported that four of his articles that were rejected from multiple highly-ranked journals are now among the twenty most-cited articles in those very journals since 2000.[6] These cases—along with the fact that interrater reliability among reviewers appears to be low—suggest that reviewers solicited by journals are not more competent than readers in the profession at large, and may, if anything, even be less competent—given that the profession later recognized the importance of papers that reviewers repeatedly missed. And indeed, in one particularly illuminating hoax study, twelve articles already published in highly-ranked psychology departments were resubmitted as 'new submissions' to the very journals that already published them (Peters and Ceci 1982). Of the nine that made it past desk-rejection, eight of the papers were *rejected* without reviewers or editors realizing that the journal had already published them—in many cases on the basis of 'serious methodological flaws'.

What explains these facts? Offhand, it seems intuitive to think that prestigious journals will select particularly accomplished reviewers, who should then be more competent than the typical reader in the discipline. How can we square this with the above anecdotes and research indicating that journal-

[6] `https://philosopherscocoon.typepad.com/blog/2015/09/stanley-on-peer-review.html`.

solicited reviewers are highly unreliable?

Although we can only speculate, there are a number of plausible reasons to think that journal-solicited reviewers are likely to be more biased than average reviewers in an academic population at large. First, suppose it is true that journals—particularly highly-selective ones—tend to commission reviewers from comparatively well-accomplished figures. In that case, the reviewers selected plausibly have particular biases—such as bias for their own views and work (as they have a vested interest in their views remaining influential), bias for particular arguments that have become influential (e.g., by other influential authors they admire or are personally acquainted with), and so on.

Second, as many have noted, the journal-solicited peer review system arguably introduces biases of its own. First, journal reviewers arguably have incentives to look for reasons to reject papers. Because they know that the journal they are reviewing for has a high rejection rate, their presumption may be that they should recommend rejection, unless they are overwhelmingly convinced the paper is excellent. This potentially lowers the risk of false positives (accepting bad papers) but increases the risk of false negatives (rejecting good papers). Notice that at least anecdotally, these incentives seem borne out by the results of traditional journal-solicited peer review process—as illustrated in Nobel Prize winning economics papers otherwise inexplicably rejected by reviewers. Second, journal editors plausibly have grounds to be strongly biased in favor of avoiding false positives as well, as publishing a bad paper may harm the journal's reputation. Finally, by explicitly selecting 'specialists' to review papers—that is, people who have already published on a submitted paper's topic—a journal-solicited peer review system runs a serious risk of *groupthink*. Indeed, consider two causal antecedents of groupthink: *group cohesiveness* and *insularity* (Janis 1972). Both are arguably embedded in journal-solicited peer review, where it appears to be standard practice for editors to seek out reviewers who are *specialists*—individuals

who have all chosen to work in a similar area, who may attend conferences together, share unpublished work among each other, and so on. Conversely, journal-solicited peer review does not appear to regularly involve a practice empirical evidence suggests serves to *prevent* groupthink: the stimulation of *intellectual conflict* (Turner and Pratkanis 1994) through the inclusion of outside perspectives (Janis 1972, pp. 209–215). Because a crowd-sourced peer review system would not only invite specialists *and* non-specialists alike to review the quality of new papers, but also potentially give reviewers the opportunity to *contest* each other's reviews (see §6.3 below), such a process would be designed to generate the kind of intellectual conflict necessary for combating groupthink. Similar arguments have been made by philosophers under the label of 'epistemic diversity': scientists actively pursuing opposing theories or methodologies is often fruitful (Feyerabend 1975, Lakatos 1978, Longino 1990, Kitcher 1993, Zollman 2010).

In sum, there are a number of plausible reasons to believe that, even if journal-solicited review processes recruit 'the best reviewers' (which we have no clear empirical evidence for), the reviewer selection process and incentive structure of journal-solicited peer review introduce biases that are likely to be less pronounced or more evenly distributed in a general population of readers in the discipline.

To be clear, readers in the general academic population will tend to have biases of their own. Like individual reviewers solicited by journals, readers in a larger population plausibly have vested-interests in advancing their own views. Some may be more concerned with avoiding false positives than false negatives, others more concerned with the converse, and so on. The point of our argument is not primarily that crowd-sourced peer reviewers are likely less biased than journal-solicited reviewers. Our argument is (1) there is no clear evidence that journal-solicited reviewers are more accurate or less biased than the reviewer pool at large; (2) there is *some* anecdotal evidence they may in fact be less accurate or more biased; but more importantly (3)

whatever biases there are in an academic population, our three jury theorems suggest that in a larger population, these biases tend to *cancel each other out* more than in a smaller population. In a smaller population, distorting biases are much more likely to lead the 'jury' (e.g., two reviewers) to the wrong verdict.

Consequently, we submit that our current evidence on the whole provides no clear support for the proposition that journal-solicited reviewers are more accurate in their judgments of paper quality than members of the profession at large. Given that the evidence is unclear, we believe it is more appropriate to assume the null hypothesis: the hypothesis that accuracy and bias *are* randomly distributed in an academic population—unless and until clear evidence is provided to the contrary.

6.3 Failures of Independence in Crowd-Sourced Peer Review?

Another worry about our argument is that the jury theorems hold only when votes are probabilistically independent. This assumption seems plausible for journal-solicited peer review: each reviewer judges a given paper without knowledge of what other reviewers think. Conversely, in a crowd-sourced peer review system, one reviewer's evaluation of a paper may influence evaluations by others—potentially generating a snowball effect (e.g., if one early influential reviewer judges a paper negatively, others may judge the paper negatively as a result). When votes in a jury are correlated, the collective competence of a jury may be *lower* than the competence of individual jurors (Kaniovski and Zaigraev 2011).

Our reply begins by noting some technical points. First, correlated votes only undermine the Condorcet theorem when the competence of individual reviewers is low and correlation between their opinions high (Kaniovski and Zaigraev 2011). Second, for the results in §4 and §5 to be undermined, an even stronger condition needs to hold: correlation between reviewer opinions

needs to *increase* systematically as the number of reviewers increases. If reviewer opinions are correlated but the correlation coefficient is constant, the first and more important part of our theorems—that the probability of a judgment close to the correct one increases with the number of reviewers—still holds true, even if the second part—that this probability goes to one in the limit—fails.

Although the research we discussed earlier on the low interrater reliability of peer review reports suggest that reviewer competence may indeed be low, we believe there are ample reasons to doubt that the correlation of votes in a crowd-sourced peer review system would be high, or that it would increase with the number of reviewers. This is due to academic training and professional incentives, which encourage academics to *evaluate* arguments and *counter* arguments they find unpersuasive. In a crowd-sourced peer review system, readers of a given paper would be encouraged to *contest* evaluations they find unpersuasive, leading to poorly correlated votes. To see how, consider the kinds of discussions that typically occur in the pages of journals after an article or book is published. To take just one famous example, consider how John Rawls' book *A Theory of Justice* was received. After its publication, some commentators were highly critical of Rawls' arguments (e.g., Hare 1973a,b, Nowell-Smith 1973, Barry 1973), others more sympathetic (e.g., Mandelbaum 1973), and so on. In time, scholars in the literature then began debating *each other*: namely, whether a particular 'flaw' in Rawls' book really is a flaw or not (e.g., Bedau 1975). After much debate, a *general consensus* eventually emerged that Rawls' work is important yet flawed in particular ways. Given that this is what philosophers (and academics more generally) are trained to do *post*-publication, there are ample reasons to believe that evaluators in a crowd-sourced peer-review model would do something similar: present their own evaluations of a given piece of work, but then *contest* other reviewers' evaluations—arguing that other reviewers are correct, mistaken, and so on. Because different readers of a given

paper may arrive at very different judgments of a paper's quality, and seek to contest other readers' evaluations they disagree with, there are reasons to think that 'votes' in a crowd-sourced model would be poorly correlated.

Moreover, we think there is good reason to think that expert reviewers in particular will form their opinions independently. Recall that in §6.1 we introduced a notion of expert reviewers (based on colleagues' endorsements) whose scores potential readers may want to consider separately as a way to guard against internet trolls. We posit that part of what it means to be an expert in a certain area is to have the ability to form one's opinion on a piece of work in one's area of expertise independently of others. The idea is that a genuine expert cannot just be somebody who happens to have more true beliefs than average about a domain. Such a person might just have good access to a genuine expert who passes on reliable information about the topic matter, but lacks an important part of scientific expertise, which has been called 'contributory expertise' in the literature (Collins et al. 2016). The contributory expert must also know the methods and heuristics one ought to adopt to reliably arrive at true beliefs about the topic (Licon 2012, p. 451). We take it that just what one's own knowledge of these methods is meant to do is make the contributory expert someone whose beliefs are relatively independent of their peers, conditional on the truth. If this is granted, at least the set of expert reviewers will satisfy the independence condition, and hence all the conditions for our jury theorems to apply. Assuming that the average number of expert reviewers under crowd-sourced peer review is at least as high as the average number of reviewers under journal-solicited peer review, it follows (at minimum) that basing one's opinion on the average expert reviewer score is better than basing one's opinion on journal-solicited peer review.

We are of the further opinion that, barring cases of internet mobs, correlation among non-expert reviewers will also tend to be low, and so basing one's opinion on the overall average score is even better than basing one's

opinion on the average expert reviewer score (as this will make for an even larger jury). But for the skeptical reader who thinks that there will be sufficient correlation among non-expert reviewers' scores such that something like the reverse of our jury theorems holds, we maintain the claim that the more conservative approach of looking only at expert reviewers' scores would still be an improvement on the status quo.

7 Conclusion

We have argued that if the presuppositions which pre-publication peer review is based upon are correct, then modified Condorcet jury theorems suggest that under those very conditions an open, crowd-sourced model of post-publication peer review would do better at directing scientists towards better work. This leaves two major questions for future research open. First, are the presuppositions of pre-publication peer-review (which underlie our argument against it) correct? The brief arguments we gave for these presuppositions here should be supplemented with more sustained, and often empirical, socio-epistemic inquiry. Second, is it actually desirable to direct scientists towards the best work? This might seem good for individual scientists but that is not yet to argue for its socio-epistemic optimality (Mayo-Wilson et al. 2011). We leave both these projects for future work.

If such positive answers can be had, we note that the practical difficulties with implementing our proposal are substantial, but not insurmountable. As we have repeatedly noted, online forums for public peer-review (such as the arXiv in physics) already exist—they even already serve as a primary point of publication in some fields—and it is not beyond our capacities to add the necessary features on some expanded version of these venues. What is presently lacking is the will. We thus hope that our paper goes towards building this will, and we are able to take further and more systematic advantage of the combined wisdom of the scientific community.

References

Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, 82(3):424–453, 2015. doi: 10.1086/681766. URL http://dx.doi.org/10.1086/681766.

J. Scott Armstrong. Combining forecasts. In J. Scott Armstrong, editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 417–440. Kluwer, New York, 2001. ISBN 0-306-47630-4.

Shahar Avin. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. doi: 10.1093/bjps/axx059. URL http://dx.doi.org/10.1093/bjps/axx059.

Brian M. Barry. *The Liberal Theory of Justice: A Critical Examination of the Principal Doctrines in a Theory of Justice by John Rawls*. Clarendon Press, Oxford, 1973.

Hugo Adam Bedau. Review of the liberal theory of justice: A critical examination of the principal doctrines in a theory of justice by John Rawls by Brian Barry. *Philosophical Review*, 84(4):598–603, 1975. doi: 10.2307/2183864. URL http://dx.doi.org/10.2307/2183864.

Lutz Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011. ISSN 1550-8382. doi: 10.1002/aris.2011.1440450112. URL http://dx.doi.org/10.1002/aris.2011.1440450112.

David Christensen. Epistemology of disagreement: The good news. *The Philosophical Review*, 116(2):187–217, 2007. doi: 10.1215/00318108-2006-035. URL http://dx.doi.org/10.1215/00318108-2006-035.

Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989. ISSN 0169-2070. doi: 10.1016/0169-2070(89)90012-5. URL http://dx.doi.org/10.1016/0169-2070(89)90012-5.

Stewart Cohen. A defense of the (almost) equal weight view. In David Christensen and Jennifer Lackey, editors, *The Epistemology of Disagreement: New Essays*, chapter 5, pages 98–119. Oxford University Press, Oxford, 2013. doi: 10.1093/acprof:oso/9780199698370.003.0006. URL http://dx.doi.org/10.1093/acprof:oso/9780199698370.003.0006.

Harry Collins, Robert Evans, and Martin Weinel. Expertise revisited, part II: Contributory expertise. *Studies in History and Philosophy of Science Part A*, 56:103–110, 2016. ISSN 0039-3681. doi: 10.1016/j.shpsa.2015.07.003. URL http://dx.doi.org/10.1016/j.shpsa.2015.07.003.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Adam Elga. Reflection and disagreement. *Noûs*, 41(3):478–502, 2007. doi: 10.1111/j.1468-0068.2007.00656.x. URL http://dx.doi.org/10.1111/j.1468-0068.2007.00656.x.

Paul Feyerabend. *Against Method*. New Left Books, London, 1975.

Joshua S. Gans and George B. Shepherd. How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, 8 (1):165–179, March 1994. doi: 10.1257/jep.8.1.165. URL http://dx.doi.org/10.1257/jep.8.1.165.

R. M. Hare. Rawls' theory of justice—I. *The Philosophical Quarterly*, 23 (91):144–155, 1973a. doi: 10.2307/2217486. URL http://dx.doi.org/10.2307/2217486.

R. M. Hare. Rawls' theory of justice—II. *The Philosophical Quarterly*, 23 (92):241–252, 1973b. doi: 10.2307/2218002. URL http://dx.doi.org/10.2307/2218002.

Stephan Hartmann and Jan Sprenger. Judgment aggregation and the problem of tracking the truth. *Synthese*, 187(1):209–221, 2012. ISSN 0039-7857. doi: 10.1007/s11229-011-0031-5. URL http://dx.doi.org/10.1007/s11229-011-0031-5.

Remco Heesen. Why the reward structure of science makes reproducibility problems inevitable. *The Journal of Philosophy*, 115(12):661–674, 2018. ISSN 0022-362X. doi: 10.5840/jphil20181151239. URL http://dx.doi.org/10.5840/jphil20181151239.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, forthcoming. ISSN 1573-0964. doi: 10.1007/s11229-016-1294-7. URL http://dx.doi.org/10.1007/s11229-016-1294-7.

Irving L. Janis. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, 1972.

Serguei Kaniovski and Alexander Zaigraev. Optimal jury design for homogeneous juries with correlated votes. *Theory and Decision*, 71(4):439–459, October 2011. ISSN 1573-7187. doi: 10.1007/s11238-009-9170-2. URL http://dx.doi.org/10.1007/s11238-009-9170-2.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993. ISBN 0195046285.

Richard L. Kravitz, Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney. Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors

care? *PLoS ONE*, 5(4):e10072, 2010. doi: 10.1371/journal.pone.0010072. URL http://dx.doi.org/10.1371/journal.pone.0010072.

Imre Lakatos. *The Methodology of Scientific Research Programmes*. Cambridge University Press, Cambridge, 1978.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Keith Lehrer and Carl Wagner. *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*, volume 24 of *Philosophical Studies Series in Philosophy*. D. Reidel, Dordrecht, 1981.

Jimmy Alfonso Licon. Sceptical thoughts on philosophical expertise. *Logos & Episteme*, 3(3):449–458, 2012. doi: 10.5840/logos-episteme20123325. URL http://dx.doi.org/10.5840/logos-episteme20123325.

Christin List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. ISSN 1467-9760. doi: 10.1111/1467-9760.00128. URL http://dx.doi.org/10.1111/1467-9760.00128.

Helen E. Longino. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, 1990. ISBN 9780691020518.

Maurice Mandelbaum. Review of a theory of justice by John Rawls. *History and Theory*, 12(2):240–250, 1973. doi: 10.2307/2504913. URL http://dx.doi.org/10.2307/2504913.

Carlo Martini and Jan Sprenger. Opinion aggregation and individual expertise. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weis-

berg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 9, pages 180–201. Oxford University Press, Oxford, 2017.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/661777.

Brian A. Nosek and Yoav Bar-Anan. Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, 23(3):217–243, 2012. doi: 10.1080/1047840X.2012.692215. URL http://dx.doi.org/10.1080/1047840X.2012.692215.

P. H. Nowell-Smith. Review symposium: I—a theory of justice? *Philosophy of the Social Sciences*, 3(4):315–329, 1973. doi: 10.1177/004839317300300403. URL http://dx.doi.org/10.1177/004839317300300403.

Cailin O'Connor and Justin Bruner. Dynamics and diversity in epistemic communities. *Erkenntnis*, 84(1):101–119, 2019. ISSN 1572-8420. doi: 10.1007/s10670-017-9950-y. URL http://dx.doi.org/10.1007/s10670-017-9950-y.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5:187–195, Jun 1982. ISSN 1469-1825. doi: 10.1017/S0140525X00011183. URL http://journals.cambridge.org/article_S0140525X00011183.

Richard Pettigrew. On the accuracy of group credences. In Tamar Szabó Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology*, volume 6, chapter 6, pages 137–160. Oxford University Press, Oxford, 2019.

Felipe Romero. Can the behavioral sciences self-correct? a social epistemic study. *Studies in History and Philosophy of Science Part A*, 60:55–69, 2016. ISSN 0039-3681. doi: 10.1016/j.shpsa.2016.10.002. URL http://www.sciencedirect.com/science/article/pii/S0039368116300802.

Daniel J. Singer. Diversity, not randomness, trumps ability. *Philosophy of Science*, 86(1):178–191, 2019. doi: 10.1086/701074. URL http://dx.doi.org/10.1086/701074.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/681768.

Marlene E. Turner and Anthony R. Pratkanis. Social identity maintenance prescriptions for preventing groupthink: Reducing identity protection and enhancing intellectual conflict. *International Journal of Conflict Management*, 5(3):254–270, 1994. doi: 10.1108/eb022746. URL http://dx.doi.org/10.1108/eb022746.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/644786.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35, 2010. ISSN 0165-0106. doi: 10.1007/s10670-009-9194-6. URL http://dx.doi.org/10.1007/s10670-009-9194-6.