Judgment Aggregation in Science

Liam Kofi Bright* Haixin Dang† Remco Heesen‡

October 31, 2014

Abstract

This paper raises the problem of judgment aggregation in science. The problem has two sides. First, how do scientists decide which propositions to assert in a collaborative document? And second, how should they make such decisions? The literature on judgment aggregation is relevant to the second question. Although little evidence is available regarding the first question, it suggests that current scientific practice is not in line with the most plausible recommendations from the judgment aggregation literature. We explore the evidence that is presently available before suggesting a number of avenues for future research on this problem.

1 Introduction

Science is an increasingly collaborative enterprise. Not only are single-authored papers now almost unheard of, massive collaborations such as the ATLAS collaboration at CERN are now common as well.

*Department of Philosophy, Baker Hall 161, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. Email: lbright@andrew.cmu.edu.

†University of Pittsburgh, Department of History and Philosophy of Science, 1017 Cathedral of Learning, Pittsburgh, PA 15260, USA. Email: had27@pitt.edu.

‡Department of Philosophy, Baker Hall 161, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. Email: rheesen@cmu.edu. The author list is alphabetical.

Widespread collaboration, for all its virtues, brings with it a new set of problems. One such problem is the topic of this paper. It is the problem of determining what to say in a co-authored talk or paper. Given a group of scientists and their opinions on various matters related to the paper, what does "the group" think?

The answer to this question has obvious epistemic and ethical consequences. If the group lets each member write their own section of the paper, the epistemic status of the statements in a given section is tied directly to its author, and ethical responsibility (say, in cases of suspected fraud) can be divided as well. Whereas if the group only publishes work that all authors have agreed to unanimously, epistemic and ethical responsibilities are divided more evenly among the group.

Despite these important consequences, the above question, which we will call the question of *judment aggregation in science*, has received very little attention. We are not aware of any scholarly work that directly addresses either the descriptive question how scientists aggregate judgments or the normative question how scientists in particular should aggregate judgments.

However, there is a literature on judgment aggregation in general that we think could fruitfully be applied to this question. It provides a formal framework for studying judgment aggregation in science and putative answers to the normative question.

In this paper we show that the problem faced by collaborating scientists has the formal structure assumed in the judgment aggregation literature (section 2). We survey this literature and extract normative recommendations (section 3). We look at what scientists say they do and what scientists say they should do, and find that they diverge from each other and from the recommendations of the judgment aggregation literature (section 4). We conclude that scientific practice could be improved by more empirical and normative investigation of judgment aggregation in science.

2 Motivation

In November 1997, the National Institutes of Health (NIH) organized a consensus conference on the topic of acupuncture. Its self-described goal was to "provide health care providers, patients, and the general public with a responsible assessment of the use and effectiveness of acupuncture for a variety of conditions" (NIH 1997, p. 1).

The conference consisted of a panel of 12 scientists with various kinds of relevant expertise, with input (through presentations) from another 25 experts. Another 1,200 people were in attendance as audience members.

Over the course of the three-day conference, the panel developed answers to the (pre-defined) questions about the role and efficacy of acupuncture as a treatment option, its biological effects, issues in integrating acupuncture in today's health-care system, and directions for future research. These answers were written up into a consensus statement, subject to revision based on expert and audience comments:

> The panel composed a draft statement, which was read in its entirety and circulated to the experts and the audience for comment. Thereafter, the panel resolved conflicting recommendations and released a revised statement at the end of the conference... The draft statement was made available on the World Wide Web immediately following its release at the conference and was updated with the panel's final revisions. (NIH 1997, p. 1)

This process is typical of consensus conferences, which have been held by the NIH and other organizations with some frequency over the last decades (Wortman et al. 1988). We draw attention to four features of this process:

1. A final document is produced which asserts certain propositions.
2. This final document purports to reflect the collective view of a group of people: at minimum, the 12 panel members; at best, the consensus of "the field". It does not purport to reflect the view of any one individual.

3. The individuals involved in the production of the document have views on the propositions asserted in it, and these serve as input to the production of the document.

4. The propositions asserted in the document are logically related (some propositions in the document are asserted to follow from other ones).

This raises a problem which we will call a “judgment aggregation problem”. A *judgment aggregation problem* asks how, given a group of individuals with views on a number of logically related propositions, the collective view of that group on those propositions is to be determined.

Consensus conferences are a particularly clear instance of a judgment aggregation problem. But this kind of problem occurs all over science. In particular, any time two or more scientists write a paper (or a talk) together, they have to solve a judgment aggregation problem (all of the above four features are present in such a case). With collaborative work now the norm in science, judgment aggregation problems are everywhere.

These considerations motivate the following two questions.

1. A descriptive question: How do scientists in fact solve judgment aggregation problems?

2. A normative question: How should scientists solve judgment aggregation problems?

Regarding the former question, very little is known. We review the (informal) evidence that exists in section 4.

The latter question has been studied in a branch of social choice theory called “judgment aggregation”. The problem was first raised in the legal context by Kornhauser and Sager (1993). Formal study of the problem took off in earnest with the work of List and Pettit (2002), which spawned a thriving literature.

This literature is concerned with judgment aggregation problems in general, not with such problems as faced in particular by collaborating scientists.

In the next section we review some of this literature with the aim of seeing whether it yields any particular recommendations for scientists.

3 Possibilities and Impossibilities in Judgment Aggregation

The judgment aggregation literature studies judgment aggregation functions, which take the judgments of a group of individuals on a number of logically related propositions as input, and yield a collective judgment on those propositions as output. It might be thought that applying this framework to collaborative scientific work is a non-starter, due to the following theorem.

Theorem 1 (List and Pettit (2002)). *There exists no judgment aggregation function which yields complete and consistent aggregate judgments and satisfies "universal domain", "anonymity", and "systematicity".*

This theorem shows that the following desiderata on a judgment aggregation function cannot be jointly satisfied.

1. Complete aggregate judgments: This means that the aggregation function yields judgments on all relevant propositions.

2. Consistent aggregated judgments: This means that the aggregation function's judgments are logically consistent.

3. Universal domain: This means that the aggregation procedure can be applied to any combination of (complete and consistent) individual judgments on the relevant propositions.

4. Anonymity: All individuals are considered equal. More formally, if the judgments of two individuals are switched, the result of the judgment aggregation procedure does not change.

5. Systematicity: All propositions are considered equal. More formally, if the judgments of all individuals are the same on two propositions, then the aggregation procedure should lead the group to either accept both of them or deny both of them.

But the theorem only presents a problem if we think they state norms that collaborating scientists should (try to) live up to. We think that anonymity, systematicity, and consistency reflect stable norms of scientific practice, for better or worse, but completeness and universal domain do not.

Anonymity reflects the norm that scientific contributions should be valued independently of the person that made them. Systematicity reflects the norm that scientists should be unbiased: their methods should not prejudge the truth or falsity of any proposition. Consistency reflects the norm that a paper (or talk, or consensus statement) should not contradict itself.

Completeness requires that the paper (or talk, or consensus statement) pronounces a judgment on all relevant propositions. Even apart from the question how to determine which propositions are relevant, this is obviously not a norm of science. It is true that writing on some topics requires saying something about related topics, but this is never so specific as to require either asserting or denying specific propositions.

Universal domain requires that a paper (or talk, or consensus statement) needs to be produced regardless of the views of the individuals involved, i.e., no matter how much they disagree. Again this does not seem to be a norm of science. In fact, it is commonly accepted practice, especially in large collaborations, for scientists to take their name off a paper if they find themselves unable to support its conclusions. This happened, for example, when the Collision Detector at Fermilab (CDF) produced evidence that some interpreted as evidence for mysterious extra muons: ghost particles that are suggestive of new physics. The results were published in an online preprint, but they were so controversial that nearly a third of the roughly 600 scientists involved refused to sign it (CDF Collaboration 2008).

If universal domain is dropped as a desideratum, the other desiderata can be satisfied. In this case proposition-wise majority voting emerges as a reasonable aggregation procedure. Proposition-wise majority voting, as the name suggests, considers each proposition individually and takes the collective judgment on that proposition to be whatever is the majority judgment among the individuals.

Dietrich and List (2010) prove a number of theorems to the effect that if the domain is restricted by requiring individuals' judgments to show certain kinds of "similarity" then proposition-wise majority voting satisfies all the other desiderata. They also prove that on any domain where majority voting yields consistent aggregate judgments, it is the unique aggregation procedure that satisfies anonymity (as above) and acceptance/rejection neutrality (which requires that if all individuals flip their judgment on a given proposition, the aggregate judgment should flip as well).

We conclude that the state-of-the-art normative recommendation from the judgment aggregation literature to collaborating scientists is to use proposition-wise majority voting whenever it yields consistent aggregation judgments, and not to produce a collective document at all when it does not. One might consider dropping completeness, however, the strongest normative recommendation from the judgment aggregation literature is obtained by dropping only universal domain.

4 What Is Known?

In section 2 we raised two questions concerning how scientists aggregate judgments.

1. A normative question: How should scientific collaborations aggregate their judgments?

2. A descriptive question: How do scientific collaborations in fact aggregate their judgments?

As far as we know, there have been no studies that investigated the former question directly. All the evidence that is available is from case studies of scientific collaborations that were not specifically interested in judgment aggregation, but where it was mentioned because it came up as a practical problem (we cite one such study below). We intend to remedy this, but that is beyond the scope of this paper.

With regard to the latter question, there are two sources of evidence. First, we could look at the literature on judgment aggregation, determine which of the various conditions considered in that literature are satisfied in the case of scientific collaborations, and obtain an answer that way. We gave a preliminary discussion of such conditions in section 3, and concluded that the judgment aggregation literature would arguably recommend proposition-wise majority voting.

Second, we could see what scientists have themselves said about the rules by which their judgments should be aggregated. Here we give two examples. First, the International Committee of Medical Journal Editors (ICMJE) speaks to this in a document with recommendations for authors involved in collaborative scientific projects. The document the ICMJE produces is especially significant, since it is recommended reading by the U.S. National Science Foundation to those who receive funding from it. The ICMJE guidelines are hence plausibly taken as authoritative for a wide range of scientific collaborations. The ICMJE requires that "[a]ll members of the group named as authors... should have full confidence in the accuracy and integrity of the work of other group authors" (ICMJE 2013, p. 3).

To us, this reads like a requirement of unanimity: every author should agree to every proposition asserted in the collaborative document. Note that this requirement is probably motivated by ethical rather than epistemic considerations, in particular related to assigning blame in cases of suspected fraud.

However, this norm is arguably too restrictive. Based on anecdotal evidence, we think that working scientists do not regard this norm as realistic and hence ignore it. This highlights an important reason for studying the processes of judgment aggregation in science, and in particular the normative question. Presently policy is unguided by theory, and hence is (and is seen to be) unmotivated, and is thus ignored by working scientists. If we are to have policy here it would be worth more explicit reflection on what the desiderata for such policy are, as well as what sort of policies would fulfill those desiderata.

For a second example, consider the case of Nobel Prize winner Carlo Rubbia. Leading a lab meeting of his research team at CERN in the 1980s, Rubbia found that people were evenly split on whether a particular proposition should be presented as well-confirmed at an upcoming conference.

> "So we're in a pretty shitty mess, aren't we?" he said. "I cannot neglect the fact that people who are working on it have more weight than people who aren't. It's also clear that we cannot run science on a majority basis." (Taubes 1986, p. 218)

Here Rubbia explicitly contradicts the claim (which we attributed to the judgment aggregation literature) that proposition-wise majority voting is an appropriate norm for judgment aggregation in science. This brings to light a further important reason for studying the process of judgment aggregation in science, this time focusing on the descriptive question. The manner in which one aggregates judgments can make an important difference to the content of what one accepts. For example, Rubbia's group ended up asserting a proposition which they would not have asserted if they had used majority rule. This shows that different aggregation procedures can lead to the same set of opinions resulting in different publications.

Social epistemologists and philosophers of science, as well as science scholars more generally, should thus be interested in the actual procedures of judgment aggregation scientists use. Perhaps especially if, as Rubbia suggests, scientists' behavior is not in line with what judgment aggregation theorists would themselves think natural or appropriate.

5 Conclusion

Scientists regularly face judgment aggregation problems in the process of collaborative research. The judgment aggregation literature makes it clear just how difficult it can be to solve those problems. Little is presently known about how scientists address these problems in practice. Our initial investigations, reported here, suggest that scientists are not solving these problems in a way that would be endorsed by the judgment aggregation literature. Neither are scientists solving these problems in a way that the explicit guidance of the ICMJE would suggest. Thus the main results of our initial investigations are negative, telling us only about what scientists are not doing.

Further research is needed to establish both what would be optimal for scientists to do, and what they are presently doing. The judgment aggregation literature suggests a promising method of investigation for resolving the normative question. Without departing from what is standard methodology in that field we can use the machinery of judgment aggregation theory to formulate abstract principles derived from what is already known about the normative structure of science. Once these are clearly expressed it will be possible to derive results about what functions would satisfy our various commitments. The descriptive question, on the other hand, raises its own methodological challenges, since there is typically no record left of the manner in which collaborative teams solve judgment aggregation problems, nor may the team members even be conscious of their own process. We hope that our raising these questions will lead to more attention being directed to solving them.

References

CDF Collaboration. Study of multi-muon events produced in p-pbar collisions at sqrt(s)=1.96 tev. Accessed online on September 30, 2014, October 2008. URL `http://arxiv.org/abs/0810.5357v1`.

Franz Dietrich and Christian List. Majority voting on restricted domains. *Journal of Economic Theory*, 145(2):512–543, 2010. ISSN 0022-0531. doi: http://dx.doi.org/10.1016/j.jet.2010.01.003. URL `http://www.sciencedirect.com/science/article/pii/S0022053110000141`.

ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. Accessed online on September 24, 2014, December 2013. URL `http://icmje.org/icmje-recommendations.pdf`.

Lewis A. Kornhauser and Lawrence G. Sager. The one and the many: Adjudication in collegial courts. *California Law Review*, 81(1):1–59, 1993. ISSN 00081221. URL `http://www.jstor.org/stable/3480783`.

Christian List and Philip Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18:89–110, April 2002. ISSN 1474-0028. URL `http://journals.cambridge.org/article_S0266267102001098`.

NIH. Acupuncture. *NIH Consensus Statement*, 15(5):1–34, November 3–5 1997. URL `http://consensus.nih.gov/1997/1997acupuncture107html.htm`.

Gary Taubes. *Nobel Dreams: Power, Deceit and the Ultimate Experiment*. Tempus Books, New York, 1986.

Paul M. Wortman, Amiram Vinokur, and Lee Sechrest. Do consensus conferences work? A process evaluation of the NIH Consensus Development Program. *Journal of Health Politics, Policy and Law*, 13(3):469–498, 1988. doi: 10.1215/03616878-13-3-469. URL `http://jhppl.dukejournals.org/content/13/3/469.abstract`.