What scientists know is not a function of what scientists know

P.D. Magnus*

June 13, 2012

This has been paper accepted for presentation at the Philosophy of Science Meeting in San Diego (November 2012) with the final version to be published in a supplementary volume of *Philosopy of Science*.

Abstract

There are two senses of 'what scientists know': An individual sense in which scientists report their own opinions, and a collective sense in which one reports the state of the discipline. The latter is what is of interest for the purpose of policy and planning. Yet an expert, although she can report the former directly (her opinion on some question), can only report her considered opinion of the latter (the community opinion on the question). Formal judgement aggregation functions offer more rigorous frameworks for assessing the community opinion. They take the individual judgements of experts as inputs and yield a collective judgement as an output. This paper argues that scientific opinion is not effectively captured by a function of this kind. In order to yield consistent results, the function must take into account the inferential relationships between different judgements. Yet the inferential relationships are themselves matters to be judged by experts involving risks which must be weighed, and the significance of the risk depends on value judgements.

In one sense, 'what scientists know' just means the claims which are the determination of our best science. Yet *science* is a collective enterprise; there

*Thanks to John Milanese, Heather Douglas, and Jon Mandle for comments on various parts of this project.

are many scientists who have individual and disparate beliefs. So 'what scientists know', in another sense, means the omnibus comprised of the epistemic state of scientist #1, the epistemic state of scientist #2, and so on for the rest of the community. The phrase is ambiguous between a collective and an individual meaning.

If we consult a scientific expert, either because we want to plan policy or just because we are curious, we are typically interested in the collective sense. We want to know what our best present science has to say about the matter. And the expert we consult can differentiate the two senses, too. She can relate what she as a particular scientists knows (what she herself thinks, where here sympathies lie in controversies, and so on), but she can also take a step back from those commitments to give her sense of what the community consensus or dominant opinion is on the same matters. If it is simply curiosity that has led us to consult an expert, this may be enough. When policy hangs on the judgement, however, we want more than just one expert's report on the state of the entire field.

This distinction between their personal commitments and the state of the field in their discipline is one that any scholar can make. If you think (as tradition has it) that only individuals can have *beliefs* in a strict sense, then take the expression 'opinion of the scientific community' as a *façon de parler*. If you think (as Lynn Hankinson Nelson does [10]) that the community rather than the individual *knows* in a strict sense, then suitably reinterpret 'what an individual knows' in terms of belief. The distinction I have in mind is neutral with respect to the metaphysics of social epistemology. The question is simply how we could use consultation with individuals to generate a composite, collective judgement.

Formal *judgement aggregation* offers rigorous frameworks which seem to provide what we want. In the abstract, it defines a function that takes individual scientists' judgements as inputs and yields collective judgement as an output. This assumes that the collective judgement of the scientific community depends on the separate individual judgments of the scientists — i.e., that *what scientists know* in the collective sense is a function of *what scientists know* in the individual sense.

Taking a recent proposal by Hartmann et al. [6][7] as an exemplar, I argue that judgement aggregation does a poor job of representing *what scientists know* in the collective sense. I survey several difficulties. The deepest stems from the fact that judgements of fact necessarily involve (perhaps implicit) value judgements. Where values and risks might be contentious, this entails that individual judgements cannot merely be inputs to a function. Judgement aggregation is not enough.

1 The majority and premise-majority rules

As a judgement aggregation procedure, one might naïvely survey scientists about factual matters and take any answer given by the majority of scientists to reflect the state of science. Of course, scientists would agree about a great many things that are simply not within their purview. Physicists would say that Sacramento is the capital of California, but that does not make it part of physics. So the survey should be confined to matters that are properly *scientific*. The survey must also include only legitimate scientists and exclude ignorant rabble. These restrictions are somewhat slippery, but let's accept them.

The naïve procedure is a simple function from individual judgements to an aggregate judgement: Return the judgement endorsed by a majority of the judges. Call this the *majority* rule.

The *majority* rule has the nice features that it treats every judge equally and that it does not bias the conclusion toward one judgement or another. Yet it suffers from what's called the *discursive dilemma*: It can lead to inconsistent collective judgements, even if all the judges considered individually have consistent beliefs. In the following schematic example, there are three judges: Alice, Bob, and Charles. Each has the consistent beliefs on the matters P, Q, and $(P\&Q)$ indicated in the table below. The *majority* rule yields the inconsistent combination of affirming P and Q but denying $(P\&Q)$.

	P	Q	$(P\&Q)$
Alice	T	F	F
Bob	F	T	F
Charles	T	T	T
majority	T	T	F

The nice features of *majority* rule seem like desiderata for a judgement aggregation rule, but avoiding the discursive dilemma is another such desideratum. A good deal of ink has been spilled specifying precisely the desiderata and proving that they are together inconsistent. However, even where it can be proven that a set of desiderata cannot be satisfied in *all* cases, they may still be jointly satisfied in some instances. The *majority* rule can lead to contradiction, it does not do so in every case. As a practical matter, we might begin by trying out a simple rule (like *majority*) and add sophistication only if the actual community has judgements like those in schematic example.[1] Even so, more sophisticated rules would be needed for corner cases.

Stephan Hartmann, Gabriella Pigozzi, and Jan Sprenger [6][7] develop a judgement aggregation rule specifically to escape the discursive dilemma. Their procedure involves polling judges only regarding matters of independent evidence. For matters which are consequences of the evidence, the procedure derives consequences from the aggregated judgements. In the simple case given in the table above, for example, the procedure would affirm P and Q (because each is affirmed by a majority) and also $P\&Q$ (because it is a consequence of P and Q). Call this the *premise-majority* rule. When it can be applied, *premise-majority* generates a consistent set of judgements.

There are several difficulties with *premise-majority*, as a way of aggregating expert scientific opinion.[2]

First, *premise-majority* inevitably produces some determinate answer. As Brams et al. [3] show, it is possible for a combination of separate elections to result in an overall outcome that would not be affirmed by any of the voters. Moreover, a judge's inconsistency will necessarily be between some belief about evidence and some belief about the consequences of the evidence — since the evidence claims are stipulated to be independent — but *premise-majority* does not query their beliefs about consequences at all. So it will generate a consistent set of judgements even if many or all judges are inconsistent. As such, *premise-majority* will generate determinate results even when the community is confused or fractured into competing camps. But, in considering scientific opinion, we certainly only want to say that there is something 'scientists know' when there is a coherent scientific community.

Second, applying the rule requires a division between the judgements that are evidence and the ones that are conclusions. As Fabrizio Cariana notes, *premise-majority* "requires us to isolate, for each issue, a distinguished set

[1] The strategy of adding complications only as necessary can be applied generally to decision problems. For example, intransitive preferences wreck dominance reasoning. Yet one might presumptively employ dominance reasoning until one actually faces a case where there are intransitive preferences.

[2] Since Hartmann et al. are thinking about the general problem of judgement aggregation, rather than the problem of expert elicitation, these are objections to the application of the rule rather than to the rule *as such*.

of logically independent premises" [4, p. 28]. He constructs a case involving three separate, contentious claims and an agreed upon constraint, such that any two of the three claims logically determines the third. It would be arbitrary to treat two of the claims as evidence (and so suitable for polling) and the third as a consequence (and so fixed by inference). The *premise-majority* rule simply is not applicable in cases where the line between premises and conclusions is so fluid. This difficulty leads Cariani to conclude only that *premise-majority* will sometimes be inapplicable; so he suggests, "Different specific aggregation problems may call for different aggregation rules" [4, p. 29]. Yet the problem is especially acute for scientific judgement, because inference can be parsed at different levels. Individual measurements like '35° at 1:07 AM' are not the sort of thing that would appear in a scientific publication; individual data points are unrepeatable and not something about which you would query the whole community. Yet they do, of course, play a rôle in inference. At the same time, scientists may take things like the constancy of the speed of light to be evidence for a theory; the evidence here is itself an inference from experiments and observations. There are different labels for these different levels. Trevor Pinch [12] calls them observations of differing *externality*. James Bogen and James Woodward [2] distinguish *data* from *phenomena*. Since we might treat the same claims as premises or conclusions, in different contexts, it is unclear what we would poll scientists about if we applied *premise-majority*.

Third, *premise-majority* is constructed for cases where the conclusion is a deductive consequence of the premises. In science, this is almost never the case.[3] Scientific inference is ampliative, and there is uncertainty not only about which evidence statements to accept but also about which inferences ought to be made on their basis. One might avoid this difficulty by including inferential relations among the evidential judgements. To take a schematic case, judges could be asked about R and $(R \to S)$; if the majority affirms both, then *premise-majority* yields an affirmative judgement for S.

One might worry that this suggestion treats ampliative, scientific inference too much like deductive consequence. The worry is that actual scientists might accept a premise of the form 'If R, then typically S' but nothing so strong as $R \to S$. It is possible for inferences based on weaker conditionals

[3] I say 'almost' because sufficiently strong background commitments can transform an ampliative inference into a deduction from phenomena. Of course, we accept equivalent inductive risk when we adopt the background commitments; cf. [9].

about what is merely typical to lead from consistent premises to inconsistent conclusions. To answer the worry, one might appeal to what John Norton [11] calls a *material theory of induction.* The central idea is that most of inductive risk in ampliative inferences is shouldered by conditional premises; Norton calls the premises *material postulates.* So — in answer to the worry — one might think that asking about material postulates would allow us to use the *premise-majority* rule to aggregate scientific judgements about many even though not absolutely all matters.

A deeper problem with the suggestion is that it presumes that scientists can say, independently of everything else, whether the inference from R to S is appropriate. That is, it assumes that material postulates can be evaluated on a ballot separately from everything else. In the remainder of the paper, I argue that this idealizes science too much. Whether a scientific inference is appropriate must be informed by *more* than just the particular evidence — the appropriate scientific conclusion depends (at least in some cases) on the risks and values involved.

In the next section, I spell out more clearly the way in which inference can be entangled with values and risk. In the subsequent section, I return to it as a problem for *premise-majority.* As we'll see, it becomes a problem for more than just Hartmann et al.'s specific proposal. It is a problem for any formal judgement aggregation rule whatsoever.

2 The James-Rudner-Douglas thesis

Here is a quick argument for the entanglement of judgement and values: There is a tension between different epistemic duties. The appropriate balance between these duties is a matter of value commitments rather than a matter of transcendent rationality. So making a judgement of fact necessarily depends on value commitments.

The argument goes back at least to William James, who puts the point this way: "*We must know the truth;* and *we must avoid error* — these are our first and great commandments as would-be knowers; but they are not two ways of stating an identical commandment, they are two separable laws" [8, p. 99]. Although James has in mind personal matters of conscience (such as religious belief), Richard Rudner makes a similar argument for scientific judgement. Rudner argues that

> the scientist must make the decision that the evidence is *suf-*

> *ficiently* strong... to warrant the acceptance of the hypothesis. Obviously our decision regarding the evidence and respecting how strong is "strong enough", is going to be a function of the *importance*, in the typically ethical sense, of making a mistake in accepting or rejecting the hypothesis. [13, p. 2]

There is not only a tension between finding truth and avoiding error, but also between risking one kind of error and risking another. Any particular test involves a trade-off between making the standards too permissive (and so mistakenly giving a positive answer) or making them too strict (and so mistakenly giving a negative answer). The former mistake is a *false positive* or *type I* error; the latter a *false negative* or *type II* error. There is an inevitable tradeoff between the risk of each mistake, and so there is a point at which the only way to reduce the risk of *both* is to collect more evidence and perform more tests. Yet the decision to do so is itself a practical as well as an epistemic decision. In any case, it leaves the realm of judgement aggregation — having more evidence would mean having different science, rather than discerning the best answer our present science has to a question. As such, values come into play. Heather Douglas puts the point this way, "Within the parameters of available resources and methods, some choices must be made, and that choice should weigh the costs of false positives versus false negatives. Weighing these costs legitimately involves social, ethical, and cognitive values" [5, p. 104].

Plotting a curve through these 19th, 20th, and 21st-century formulations, call this the *James-Rudner-Douglas* or *JRD thesis*: Anytime a scientist announces a judgement of fact, they are making a tradeoff between the risk of different kinds of error. This balancing act depends on the costs of each kind of error, so scientific judgement involves assessments of the value of different outcomes.

The standard objection to the thesis is that responsible scientists should not be making categorical judgements. They should never simply announce 'P' (the objection says) but instead should say things like 'The available evidence justifies $x\%$ confidence in P.' This response fails to undercut the thesis, because procedures for assigning confidence levels also involve a balance between different kinds of risk. This is clearest if the confidence is given as an interval, like $x \pm e\%$. Error can be avoided, at the cost of precision, by making e very large. Yet a tremendous interval, although safe, is tantamount to no answer at all.

Eric Winsberg and Justin Biddle [1] give a substantially more subtle reply to the standard objection. Regarding the specific case of climate modeling, Winsberg and Biddle show that scientists' estimates both of particular quantities and of confidence intervals depend on the histories of their models. For example, the results are different if scientists model ocean dynamics and then add a module for ice formation rather than vice-versa. The history of a model reflects decisions about what was considered to be important enough to model first, and so it depends on prior value judgements.

But why should the JRD thesis have consequences for expert elicitation? After all, James does not apply it to empirical scientific matters. He is concerned with religious and personal matters, and he concludes merely that we should "respect one another's mental freedom" [8, p. 109]. He does not apply it at all scientific matters where there is a community of legitimate experts.

Rudner, who does apply the thesis to empirical judgements, nevertheless hopes that the requisite values might themselves be objective. What we need, he concludes, is "a science of ethics" [13, p. 6]. Rudner calls this a "task of stupendous magnitude" [13, p. 6], but he is too optimistic. Searching for an objective ethics in order to resolve the weight of values and risks is a fool's errand. A regress would ensue: The judgements of ethical science would need to be informed by the ethically correct values so as to properly balance inductive risks, but assurance that we have the correct values would only be available as the product of ethical science. One might invoke pragmatism and reflective equilibrium, but such invocations would not give Rudner final or utterly objective values. If responsible judgement aggregation were to wait on an utterly objective, scientific ethics, then it would wait forever.

Douglas accepts that the thesis matters for expert elicitation. So she considers the concrete question of how to determine the importance of the relevant dangers. She argues for an *analytic-deliberative process* which would include both scientists and stakeholders [5, ch. 8]. Such a process is required when the scientific question has a bearing on public policy, and there are further conditions which must obtain in order for such processes to be successful. For one, "policymakers [must be] fully committed to taking seriously the public input and advice they receive and to be guided by the results of such deliberation" [5, p. 166]. For another, the public must be "engaged and manageable in size, so that stakeholders can be identified and involved" [5, p. 166]. Where there are too many stakeholders and scientists for direct interaction, there can still be vigorous public examination of the values involved. Rather than pretending that there is any all-purpose procedure, Douglas calls for "experiment with social mechanisms to achieve a robust dialog and potential consensus about values" [5, p. 169]. Where consensus is impossible, we can still try to elucidate and narrow the range of options. Douglas' approach is both a matter of policy (trying to increase trust in science, rather than alienating policymakers and stakeholders) and a matter of normative politics (claiming that stakeholders' values are ones that scientists should take into consideration). In cases where these concerns are salient, saying *what scientists know* will depend on more than just the prior isolated judgements of scientists — but moreover on facts about the actual communities of scientists, policymakers, and stakeholders.

Arguably, Douglas' concerns will not be salient in all cases. Some science is far removed from questions of policy. So the significance of the JRD thesis may depend on the question being asked.

3 Our fallible selves

I argued above that the *premise-majority* rule was inapplicable in many scientific contexts because it only worked for cases of deductive consequence. Formally, this worry could be resolved by asking scientists about which inferences would be justified; we poll them about claims like $(E \rightarrow H)$ at the same time as we poll them about E. The JRD thesis undercuts this formal trick. Where the judgement has consequences, the inference itself is an action under uncertainty. So the appropriate inference depends on the values at stake. Schematically, whether one should assent to $(E \rightarrow H)$ depends on the risks involved in inferring H from E. Concretely, questions of science that matter for policy are not entirely separable from questions of the policy implications.

If we merely poll scientists, then we will be accepting whatever judgements accord with their unstated values. We instead want the procedure to reflect the *right* values, which in a democratic society means including communities effected by the science. Importantly, this does not mean that stakeholders get to decide matters of fact themselves; they merely help determine how the risks involved in reaching a judgement should be weighed. Nor does it mean that politicized scientific questions should be answered by political means; climate scientists can confidently identify general trends and connections, even allowing for disagreement about the values involved. What it does mean is that scientists cannot provide an account that is value-neutral in all its precise details.[4]

This is fatal to *premise-majority* as a method of determining what scientists know collectively. Moreover, it is fatal to any judgement aggregation rule that treats judges merely as separate inputs to an algorithm. The problem extends to practical policies of expert elicitation, insofar as they are procedures for enacting judgement aggregation rules. Where there are important values at stake that scientists are not taking into account or where the value commitments of scientists are different than those of stakeholders, the present judgements of individual scientists can not just be taken as givens.

An analytic-deliberative process is required, but the appropriate mechanisms are not ones which we can derive *a priori*. As Douglas argues, we need to experiment with different possibilities [5, p. 169, cited above]. There is not likely to be one universally applicable process. It will depend on facts about the communities involved. Moreover, the inference from social experiments in deliberation will itself be an inductive inference about a question that effects policy. So the inference depends importantly on value judgements about the inductive risks involved, and that means an analytic-deliberative process will be required. It would be a mistake to hope, in parallel with Rudner's appeal to a science of ethics, for an objective set of procedural norms. How best to resolve meta-level judgement about experiments in social arrangements is as much a contingent matter as how to socially arrange object-level expert consultation. We start with the best processes we can muster up now, and we try to improve them going forward. Minimally, however, we can say that future improvements should not elide the rôle of values, as formal judgment aggregation functions do, but explicitly accommodate it.

References

[1] Justin Biddle and Eric Winsberg. Value judgements and the estimation of uncertainty in climate modelling. In P.D. Magnus and Jacob Busch, editors, *New Waves in Philosophy of Science*, pages 172–197. Palgrave MacMillan, Basingstoke, Hampshire, 2010.

[4] Douglas [5, esp. ch. 6] provides an excellent discussion of how (what I have called) the JRD thesis is compatible with objectivity.

[2] James Bogen and James Woodward. Saving the phenomena. *Philosophy of Science*, 97(3):303–352, July 1988.

[3] Steven J. Brams, D. Marc Kilgour, and William S. Zwicker. The paradox of multiple elections. *Social Choice and Welfare*, 15(2):211–236, 1998.

[4] Fabrizio Cariani. Judgment aggregation. *Philosophy Compass*, 6(1):22–32, 2011. doi:10.1111/j.1747-9991.2010.00366.x.

[5] Heather E. Douglas. *Science, Policy, and the Value-free Ideal*. University of Pittsburgh Press, 2009.

[6] Stephan Hartmann, Gabriella Pigozzi, and Jan Sprenger. Reliable methods of judgment aggregation. *Journal for Logic and Computation*, 20:603–617, 2010.

[7] Stephan Hartmann and Jan Sprenger. Judgment aggregation and the problem of tracking the truth. *Synthese*, forthcoming.

[8] William James. The will to believe. In Alburey Castell, editor, *Essays in Pragmatism*, pages 88–109. Hafner Publishing Co., New York, 1948. Originally published June, 1896.

[9] P.D. Magnus. Demonstrative induction and the skeleton of inference. *International Studies in the Philosophy of Science*, 22(3):303–315, October 2008. doi:10.1080/02698590802567373.

[10] Lynn Hankinson Nelson. *Who Knows: From Quine to a Feminist Empiricism*. Temple University Press, Philadelphia, 1990.

[11] John D. Norton. A material theory of induction. *Philosophy of Science*, 70(4):647–670, October 2003.

[12] Trevor Pinch. Towards an analysis of scientific observation: The externality and evidential significance of observational reports in physics. *Social Studies of Science*, 15:3–36, 1985.

[13] Richard Rudner. The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1):1–6, January 1953.