On the Formal Properties of Weighted Averaging as a Method of Aggregation
Author(s): Carl Wagner
Source: *Synthese*, Vol. 62, No. 1, Consensus (Jan., 1985), pp. 97-108
Published by: Springer
Stable URL: http://www.jstor.org/stable/20116086

CARL WAGNER

ON THE FORMAL PROPERTIES OF WEIGHTED AVERAGING AS A METHOD OF AGGREGATION

I first encountered Keith Lehrer's work on consensus during the summer of 1977,[1] and was immediately intrigued by the possibility of developing a formal account of his model of rational group decisionmaking. I adopted as a model for this enterprise the axiomatic method of social choice theory and was not surprised to discover that the question of how to aggregate probabilities was as complex and problematic as the question of how to aggregate preferences and utilities. The former issue has only recently received the kind of attention which has been directed at the latter for several decades, and is finally beginning to be as vigorously debated, as attested to by the essays in this volume. I am grateful to Barry Loewer for organizing this forum on *Rational Consensus in Science and Society*. My replies to the preceding critical essays follow.

1. DAVIS BAIRD

I agree with Baird that averages may conceal differences in standard deviations as well as in other distributional aspects of a set of numerical estimates. In certain decisionmaking situations one is well advised to pay attention to such differences and to eschew a premature compromise in the form of an average. In other situations it is not so clear that distributional information is useful. Suppose, for example, that I am trying to decide whether to undergo a particular surgical procedure and consult a group of experts on the probability that the operation will succeed. Assume that I must decide whether or not to undergo this operation based solely on their advice. (Imagine that circumstances rule out delaying the decision in order to give these experts time for further research, and that there is no more qualified group to whom I can turn for advice.) Suppose that they report to me a consensual probability of success equal to 0.6, based on a consensual weighted average of their initial estimates. It seems to me that I ought to use this number in calculating the expected utility of undergoing the operation, irrespective of the distributional aspects of their initial set of estimates.

Synthese **62** (1985) 97–108. 0039–7857/85/0621–0097 $01.20

Baird may reply that the scenarios in which (a) they all initially agree that 0.6 is the probability of success, (b) their initial opinions are widely scattered about 0.6, and (c) their initial opinions are concentrated in two "camps" on either side of 0.6 represent very different sorts of situations. I agree that one can make interesting conjectures about the state of the art involved in making the prediction in question, depending on which of these scenarios obtains, but it seems to me that such speculations are irrelevant to my decision. Perhaps Baird would reply that the knowledge that scenario (b) obtained prior to the group's coming to consensus should incline me to caution and that I should employ in my calculations of expected utility the smallest initial estimate of the probability in question. Even if the individual responsible for that estimate received the lowest consensual weight of any of the surgeons in the group? This is not caution, but pessimism, and is no more warranted than embracing the largest initial estimate of the probability in question.

Similarly, suppose that scenario (c) obtained prior to the group's coming to consensus. On what grounds should I adopt the prior estimate of one camp in preference to that of the other? Remember that I have no independent means of judging the relative expertise of the two groups.

In any case, the distributional information which Baird wishes to accord more prominence relates to the *initial* set of estimates of the probability in question. The crucial issue is not how these estimates are distributed, but what, if any, the final consensual point estimate is. Nothing in our method forces individuals to grant weight to the opinions of others. How should I make my decision, if not on the basis of an uncoerced consensus of the experts?

If my scheme of utilities is such that a consensual estimated probability of success equal to 0.6 results in my choosing to undergo the operation, I may, of course, enter the operating room in somewhat different psychological states, depending on the prior distribution of estimates of the probability in question. I may choose to be encouraged by the fact that one or more surgeons initially estimated the probability of success to be 0.8 or allow myself to be plagued by apprehension about the fact that some of them initially estimated this probability to be 0.4 (in which case it might be best for me *not* to be apprised of this distributional information, since anxiety can have physiological consequences). But such emotional distractions are endemic to decisionmaking. Indeed, the normative model which endorses decisionmaking on the basis of expected utilities may be viewed in part as an attempt to counteract the retrograde influence of such considerations. Baird, albeit unwittingly, may be contributing to bringing them to the fore.

Baird also notes that in, say, a 2-person decision problem, the case in which each person assigns the other the same small weight and the case in which each person assigns the other the same substantial weight both yield the consensual weight vector $(\frac{1}{2}, \frac{1}{2})$. His conclusion is that our method fails to differentiate mutual low regard from mutual high regard. He is of course assuming the applicability of our elementary model, so that the weights in question are granted across a hierarchy of evaluative skills. Viewed from this perspective, his example seems considerably less disturbing than at first glance. May I not come to accord another's opinion the same weight as my own by different paths? A sequence of measured, even grudging, acts of concession may, after all, cumulate in the same result as an immediate appreciation of another's point of view.

In both of the foregoing examples the issue between Baird and ourselves is whether we have ignored differences that make a difference. The differences which he wishes to emphasize seem to me to be data for psychologists rather than decision theorists.

2. ROBERT LADDAGA AND BARRY LOEWER

Supposing that individuals $i = 1, \ldots, n$ have assigned subjective probabilities $p_i(s_j)$ to a sequence of pairwise contradictory, exhaustive propositions s_j, $j = 1, \ldots, k$, we have proposed aggregating their assignments into a single consensual assignment p by a rule of the form

$$(2.1) \qquad p(s_j) = \sum_{i=1}^{n} w_i p_i(s_j).$$

The function p is extended to disjunctions $u = s_{j_1} \vee \cdots \vee s_{j_r}$ of atomic propositions by the natural and obvious rule

$$(2.2) \qquad p(u) = \sum_{k=1}^{r} p(s_{j_k}),$$

and consensual conditional probabilities for disjunctions u and t are to be computed by the rule

(2.3) $p(u \mid t) = p(u \wedge t)/p(t)$.

Now it might occur to one to try alternatively to calculate the consensual conditional probability $p(u \mid t)$ as a weighted average of the individual conditional probabilities $p_i(u \mid t)$. Laddaga and Loewer note that the weights w_i, used to aggregate unconditional probabilities, are not suitable for this task since, in general,

$$(2.4) \qquad p(u \mid t) \neq \sum_{i=1}^{n} w_i p_i(u \mid t).$$

This is not a proof of some grave defect in our method, as Laddaga and Loewer seem to think, but just an indication of the fact that the aggregation of conditional probabilities involves subtler considerations than the aggregation of unconditional probabilities. The correct way to aggregate conditional probabilities, as is well known (Raiffa, 1968, Chapter 8, §11), is given by the weighted average

$$(2.5) \qquad \sum_{j=1}^{n} \beta_i p_i(u \mid t),$$

where

$$(2.6) \qquad \beta_j = w_i p_i(t) \Big/ \sum_{i=1}^{n} w_i p_i(t).$$

It is easy to check that (2.5) yields the same result as (2.3), a fact which Levi (1980) puts to crucial use in his theorem on conditionalization and convex sets of probability distributions.

It is obvious that under aggregation by arithmetic averaging, individuals may assign probabilities to the atomic propositions $s_1, \ldots, s_k$ in such a way that some pair of disjunctions, u and t, of these propositions turn out to be independent for each of their assignments, but not independent for the consensual assignments. I have argued elsewhere at length (Lehrer and Wagner, 1983) that this is a bad thing only if independence is of epistemic significance in the probability assessment problem at hand, and that in problems where the initial acts of assessment are directed at assigning probabilities to a sequence of atomic propositions, independence is an after-the-fact formal curiosity.

A further example may be in order. Suppose that you observe r successes in n trials and I observe r' successes in n' (different) trials, with the data generating process agreed by both of us to be a Bernoulli process. Based on our respective data I would select as the density function for the parameter p of this process

$$\beta(n, r) = (n-1)!x^{r-1}(1-x)^{n-r-1}/(r-1)!(n-r-1)!, \qquad 0 \le x \le 1,$$

and you would select $\beta(n', r')$. The appropriate consensual density function here is not a weighted average of $\beta(n, r)$ and $\beta(n', r')$, but rather $\beta(n+n', r+r')$. I cite this example both to emphasize that we have not endorsed weighted averaging as a method of aggregating density functions and also to show that the perfectly unexceptionable consensual density function $\beta(n+n', r+r')$ is not independence preserving. For if we each separately observe 1 success in 2 trials, we will each separately select $\beta(2, 1)$, the uniform distribution on $[0, 1]$, as the density function of p. For the uniform density, the propositions $p \in [1/4, 3/4]$ and $p \in [5/8, 7/8]$ are independent. This is not the case for the consensual density $\beta(4, 2) = 6x(1-x)$. Are Laddaga and Loewer prepared to argue that one should be disturbed by this failure to preserve independence?

There are of course other kinds of assessment problems where independence is epistemically significant. Such problems typically involve a sequence of independent repeated trials. Suppose, for example, that you and I are asked to assign probabilities to the 4 possible outcomes of flipping a (possibly biased) coin twice. I have information that the coin showed heads 3 times in 10 flips and you have information that coin showed heads 8 times in 20 (different) flips. Based on my sample alone I would estimate the probabilities of HH, HT, TH, and TT, respectively, as $(0.3)^2$, $(0.3)(0.7)$, $(0.7)(0.3)$, and $(0.7)^2$ and based on your sample alone, you would estimate these probabilities, respectively, as $(0.4)^2$, $(0.4)(0.6)$, $(0.6)(0.4)$, and $(0.6)^2$. In this situation it would be ridiculous to take a weighted average of our estimates of the probabilities of these events in order to produce a consensual distribution. (For some reason, Laddaga and Loewer think that our method commits us to this sort of thoughtless averaging of any probability distributions that cross our path. It doesn't.) What we should do, clearly, is to combine our samples, producing the consensual probability 11/30 of heads on any toss, and then invoke independence of the outcomes on each toss (on which we agree) to produce the consensual assignments $(11/30)^2$, $(11/30)(19/30)$, $(19/30)(11/30)$, and

$(19/30)^2$. Although one hardly notices it in such a simple case, weighted averaging is used here, not to combine our estimates of the probabilities of the events HH, HT, TH, and TT, but to combine our estimates of the probabilities of the possible outcomes $\{H, T\}$ on any toss. For combining our samples is equivalent to combining our estimates of these basic probabilities, with weights proportional to the sample sizes on which we based these estimates: $(1/3)(0.3)+(2/3)(0.4)=11/30$ and $(1/3)(0.7)+(2/3)(0.6)=19/30$.

The above example reveals a simple and generally reliable indicator of the epistemic significance of independence: When individuals all make use of independence in order to assign probabilities, i.e., when they actually calculate certain probabilities by multiplying other probabilities, they may be assumed to have a prior theoretical commitment to such independence. Failure to incorporate such instances of independence in a consensual aggregation of their individual distributions would thus ignore an epistemically significant feature of their distributions. We have seen in the above example, however, that one can easily take account of epistemically significant cases of independence within the framework of our approach to consensus. The example also shows how our method can easily accommodate a consensus that some random variable is binomially distributed, contrary to the claim of Laddaga and Loewer.

One further remark about the preservation of independence is in order. Laddaga and Loewer make the strong claim that a method of aggregating probabilities must *always* preserve independence (even in cases where it is of no epistemic significance) in order to avoid inconsistency. But the sequence of propositions which they employ to argue for this assertion are bi-level propositions which do not admit probability assignments, however they may seek to disguise this fact by their choice of notation. What they label $H \,\&\, K \,\&\, I$ is, after all, the assertion $H \,\&\, K \,\&\, P(H \,\&\, K)=P(H)P(K)$. Since the first two propositions are amenable only to first order probability assignments and the third is amenable only to a second order probability assignment (assuming one believes in higher order probabilities), the entire compound proposition can be assigned no meaningful probability.

Indeed, if Laddaga and Loewer's gambit worked, it could easily be modified to show that aggregation methods must preserve *every* feature common to all individual distributions, no matter how esoteric. Thus, for example, nonindependence and nonequality of probabilities assigned to two events would have to be preserved in consensual distributions if these properties obtained in every individual distribution. To take an even wilder example, suppose that Lehrer assigns the propositions A and B the probabilities 1/5 and 4/5 and I assign them the probabilities 2/5 and 3/5. Let F be the assertion that the probabilities of A and B are rational numbers which, in lowest terms, have a denominator equal to 5. According to Laddaga and Loewer's argument, our probability matrix for the expanded set of propositions $A \& F$, $A \& \bar{F}$, $B \& F$, and $B \& \bar{F}$ would be

	$A \& F$	$A \& \bar{F}$	$B \& F$	$B \& \bar{F}$
(L)	1/5	0	4/5	0
(W)	2/5	0	3/5	0

Now assuming that any reasonable method would assign the second and fourth propositions a consensual probability of zero, consensual probabilities for the above four events must take the form p, 0, $1-p$, and 0. Thus the consensual probability of F equals 1 and any p not of the form $k/5$, for $1 < k < 5$, would be inconsistent with this fact. But this conclusion is absurd. If for example, Lehrer and I based our initial probabilities on separate examinations of disjoint samples of 5, he detecting the presence of A in 1 case and I in 2 cases, etc., it would be eminently reasonable to assign A and B the consensual probabilities 3/10 and 7/10.

Laddaga and Loewer conclude §III of their paper with an example of a decision problem where, prior to aggregating probabilities, two individuals agree on the best course of action, and afterwards they disagree. To our (correctly) anticipated reply that that prior agreement is shown to be ill founded by the consensual aggregated probabilities, they reply that consensual probabilities based on weighted averaging are not rationally compelling. Not ever? Suppose A has seen s_1 happen 6 times each following d_1 and d_2 and s_2 4 times each. B, on the other hand, has seen (in a disjoint sample) s_1 happen 4 times each following d_1 and d_2 and s_2 6 times each. They aggregate their probabilities using the consensual weight vector $(\frac{1}{2}, \frac{1}{2})$ based on the fact that they have observed equal size samples. This results in their assigning each of the outcomes s_1 and s_2 equal consensual probability. Such consensual probabilities strike me as an impeccable basis for calculating new expected utilities. If employing them changes a prior agreement on the best course of action, so be it. We are after informed agreement, not agreement at any price.

3. ISAAC LEVI

Levi endorses, with some qualifications, Laddaga and Loewer's criticism of our method's failure to preserve independence. He asserts that representing consensus as the convex hull of a set of competing probability vectors avoids this problem. Suppose, for example, that Levi and I assign the respective probabilities $(4/9, 2/9, 1/9, 2/9)$ and $(1/9, 2/9, 4/9, 2/9)$ to a sequence of pairwise contradictory, exhaustive propositions s_1, s_2, s_3, s_4. On each of our assignments the propositions $u = s_1 \vee s_2$ and $t = s_2 \vee s_3$ turn out to be independent. If we come to agree that t is true and condition on this evidence, our respective conditional probability assignments will be $(0, 2/3, 1/3, 0)$ and $(0, 1/3, 2/3, 0)$. Levi would emphasize that both before and after conditionalization the maximum probability assigned to u by either of us is $2/3$ and the minimum probability is $1/3$. This is not preservation of independence in the sense demanded by Laddaga and Loewer, but, rather, preservation of interval estimates under conditionalization relative to an independent (i.e., irrelevant) item of evidence.

What Laddaga and Loewer demand is that any consensual assignment of *unconditional* probabilities to the propositions s_1, s_2, s_3, s_4 should be such that the propositions u and t turn out to be independent. To the extent that Levi regards each weighted arithmetic mean of our original assignments as a "potential resolution of the conflict" between our original assignments, he is potentially violating preservation of independence in the sense of Laddaga and Loewer. Indeed, the only actual resolution of our conflict which preserves independence in this sense would be for me to adopt Levi's original distribution or for him to adopt mine (see Lehrer and Wagner, 1983). To be sure, no actual violation of preservation of independence will occur under Levi's scheme if our disagreement is not resolved, but if this is virtue, it is virtue by default.

In any case, as I have argued in my reply to Laddaga and Loewer, independence is rarely of epistemic significance where, as in our model, the initial acts of assessment are directed at the probabilities of a set of pairwise contradictory, exhaustive propositions. Thus, neither the potential violation of preservation of independence by Levi's model, nor its actual violation by our model is cause for concern.

One very positive aspect of Levi's conception of consensus is its tendency to restrain premature compromise. Particularly in the area of group decisionmaking, calculating expected utilities relative to all of the probability distributions in the convex hull of a set of competing distributions seems to me to be a salutary exercise. Of course, these rival ways of evaluating two courses of action will rarely unanimously establish the superiority of one action over another.

Indeed, one course of action will decisively beat another in this sense just when the smallest expected utility attributed by anyone (using his initial subjective probabilities) to the former exceeds the largest expected utility attributed by anyone to the latter. In some cases requiring a decision to adopt one course of action over another to be supported in this way may be excessively demanding. But on issues like nuclear safety, as Levi (1980) has persuasively argued, an extreme modesty regarding our probability estimates is entirely appropriate.

4. HANNU NURMI

Most of Nurmi's critique is directed at alleged deficiencies of the method of weighted averaging as a social choice mechanism. He observes first that our method does not always pick the Condorcet winner or exclude the Condorcet loser. Condorcet criteria may be reasonable restrictions on social welfare *functions* (methods of producing a consensual ordering from a profile of individual orderings) since the inputs contain no information about depth of preference. They are not reasonable restrictions on social welfare *functionals* (methods of producing a consensual ordering from a profile of individual utilities). We have been very careful to restrict our advocacy of weighted arithmetic averaging to cases where differences in utilities are interpersonally comparable (*Consensus*, p. 118). In such cases, formal results (*Consensus*, Theorem 6.9) show that weighted arithmetic averaging is the precisely appropriate method of aggregation. Informally it is already clear, however, that if differences are interpersonally comparable, then it is not unreasonable to take them into account. The aforementioned theorem strengthens this to the assertion that they *must* be taken into account. Thus, A's much greater preference for x over y can justifiably defeat the more modest preferences of B and C for y over x, even when A, B, and C all receive equal weight.

Nurmi observes next that our method violates the weak axiom of revealed preference as well as path-independence. The examples used to show this involve the aggregation of von Neumann–Morgenstern utilities, as Nurmi forthrightly indicates. But differences in such utilities are not interpersonally comparable, and we specifically forbid the aggregation of such utilities by weighted arithmetic averaging, except in the dictatorial case where one person receives all the weight (*Consensus*, pp. 119–120).

Nurmi observes further that, under our method, it is possible that x is ranked highest for a given assignment of utilities, but loses first place for another assignment of utilities which differs from the first only in that one individual assigns x an even higher utility than he did before. This is only possible, of course, if different weights are used in the two cases. I see nothing remarkable or disturbing about this. Weights make a difference. That's what they're designed to do.

In addition, Nurmi remarks that, under our method, two groups may separately rank x highest, but x may lose first place when the groups are amalgamated, even though no one changes his utility assignments. The explanation for this lies in a change in the weights. Again, weights make a difference. In Nurmi's example individual 1, who prefers x_1, reduces his own weight and assigns the difference to individual 4, who prefers x_2. Similarly, individual 3, who prefers x_1, reduces his own weight and assigns the difference to individual 2, who prefers x_2. Individuals 2 and 4 do not, however, reciprocate. Why is it surprising or "inconsistent," as Nurmi would have it, that x_2 might now be ranked first?

In addition to the foregoing criticisms of our method as a social choice mechanism, Nurmi identifies what he takes to be a defect in the general method of arriving at consensual weights by iterated weighted averaging, i.e., by repeated multiplication of weight matrices. His observation is that a modest shift in the weights assigned by a single individual can produce a substantial shift in consensual weights, as illustrated by the matrices

$$W_1 = \begin{bmatrix} 0.50 & 0.45 & 0.05 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.85 & 0.05 & 0.00 & 0.05 & 0.00 \\ 0.05 & 0.50 & 0.45 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.50 & 0.40 & 0.10 \\ 0.05 & 0.00 & 0.00 & 0.05 & 0.80 & 0.10 \\ 0.00 & 0.00 & 0.00 & 0.10 & 0.40 & 0.50 \end{bmatrix}$$

and W_2, which is identical with W_1 except that its fifth row is 0.10 0.00 0.00 0.05 0.80 0.05. The corresponding consensual weight vectors are

$$C_1 = [0.078 \quad 0.371 \quad 0.041 \quad 0.054 \quad 0.371 \quad 0.085]$$

and

$$C_2 = [0.108 \quad 0.514 \quad 0.057 \quad 0.032 \quad 0.257 \quad 0.032].$$

This is an interesting example and Nurmi is right in demanding an explanation. The explanation is two-fold. First, the proper way to compare weights is in terms of ratios, rather than differences (see *Consensus*, pp. 131–132, for examples of how apparently substantial shifts in a weight matrix produce no change in the consensual weight vector because ratios of off-diagonal elements are unchanged; also the example at the top of p. 134). Second, it is important to keep in mind that Nurmi is assuming the applicability of our elementary model, where different weight matrices at different levels are precluded. Thus he is not simply shifting a first order weight matrix in the manner indicated above, but all higher order matrices as well. (See also in this connection my reply to Baird.) Combining the above remarks, I would describe Nurmi's example as follows: "Individual #5 doubles the weight which he assigns to #1 and halves the the weight which he assigns to #6 in each of an infinite hierarchy of weight matrices. Individual #1, on the other hand, assigns nearly half of his available weight to #2 at all levels, and no weight at all to #5. Conversely, individual #6 assigns nearly half of his available weight to #5 at all levels, and no weight at all to #2. The result is that individual #5, who previously received a consensual weight equal to that of #2, now receives only half of the consensual weight received by #2." This does not strike me as being counter-intuitive. If I double, at all levels, the weight which I give to someone who has no respect for me, and halve the weight which I give to someone with substantial respect for me, is it any wonder that I find my position undermined?

5. FREDERICK SCHMITT

Schmitt objects to the fact that we do not allow an individual to receive different weights on different competing propositions in probability assessment problems. From the observation that an individual may be more or less informed about one proposition than about another, he infers that "variant" weights should be allowed. But probabilities are not assigned to a set of mutually exclusive, exhaustive propositions in a sequence of isolated acts. The enterprise is a unified one. In order to assess the probability that Native Dancer will win the race, we need to know who his competitors will be. A formal result (*Consensus*, Theorem 6.4) supports this intuition by granting the possibility of aggregating probabilities assigned to different propositions by different functions (Axiom IA) but then showing that this apparent flexibility is precluded, largely because consensual probabilities (of a set of exclusive, exhaustive propositions) must sum to one. Schmitt is exercised to the point of exclamation ("But the rejection of this assumption is of a piece with the rejection of invariance! Consensual probabilities must obviously not sum to one . . .") by the fact that this basic property of probabilities is used in the proof. The reasons for his agitation escape me.

NOTE

[1] I was at the time a participant in the Institute on Freedom and Causality directed by Lehrer at the Center for Advanced Study in the Behavioral Sciences, which subsequently supported our work by providing me with a Fellowship during 1978–79.

REFERENCES

Lehrer, Keith and Wagner, Carl: 1981, *Rational Consensus in Science and Society*, Reidel, Dordrecht–Boston.

Lehrer, Keith and Wagner, Carl: 1983, 'Probability Amalgamation and the Independence Issue: A Reply to Laddaga', *Synthese* **55**, 339–346.

Levi, Isaac: 1980, *The Enterprise of Knowledge*, MIT Press, Cambridge, Mass.

Raiffa, Howard: 1968, *Decision Analysis*, Addison-Wesley, Reading, Mass.

Wagner, Carl: 1984, 'Aggregating Subjective Probabilities: Some Limitative Theorems', *Notre Dame Journal of Formal Logic* **25**, 233–240.

Department of Mathematics
University of Tennessee
Knoxville, TN 37996-1300
U.S.A.