Scientific Expertise and Risk Aggregation

Thomas Boyer-Kassem*†

When scientists are asked to give expert advice on risk-related questions, such as the authorization of medical drugs, deliberation often does not eliminate all disagreements. I propose to model these remaining discrepancies as differences in risk assessments and/or in risk acceptability thresholds. The normative question I consider, then, is how the individual expert views should best be aggregated. I discuss what "best" could mean, with an eye to some robustness considerations. I argue that the majority rule, which is currently often used in expert panels, has significant drawbacks.

1. Introduction. Regularly, institutions ask scientists to give expert advice on pressing risk-related questions. For instance, governmental agencies that regulate medicines regularly resort to expert panels, and national scientific academies give advice to governments and legislatures. Even after exchanging information and arguments, scientific experts often do not agree on what answer to give, and even when they do, they may disagree on the justification for this answer. How should binary decisions that involve risk assessments be made and justified within scientific expert panels? This is the central question studied in this article. As a matter of fact, many expert panels are required to make decisions by majority rule, including the advisory committees for the European and American agencies that are responsible for improving medicine, the European Medicines Agency (EMA) and the Food and Drug Administration (FDA), respectively (Hauray and Urfalino 2007; Urfalino and Costa 2015), the European Environment Agency (EEA), the European Food Safety Agency (EFSA), the European Chemicals Agency (ECHA),

Received February 2017; revised October 2017.

*To contact the author, please write to: MAPP, Université de Poitiers, 36 rue de la Chaîne, 86000 Poitiers, France; e-mail: thomas.boyer-kassem@univ-poitiers.fr.

†Many thanks to Jan Sprenger, Cyrille Imbert, Camille Aron, Franz Dietrich, Naftali Weinberger, two anonymous referees for *Philosophy of Science*, participants at the DRI seminar (ENS, Paris), the EPS seminar (TiLPS, Tilburg University), the EEN2016 conference (Paris), and the PSA2016 conference (Atlanta) for valuable comments and insights. This work has benefited from an AXA postdoctoral fellowship.

Philosophy of Science, 86 (January 2019) pp. 124–144. 0031-8248/2019/8601-0006$10.00

the International Accounting Standards Board (IASB), and scientific committees advising the European Commission.[1] But is majority rule the best decision rule?

Suppose the committee is asked whether a food additive is safe enough to be granted market authorization. The members are instructed as follows: "Grant the authorization provided that the risk that the product is harmful for human health is lower than 5%" (on some appropriate scale). The committee is composed of five experts: four of them estimate the risk associated with the food additive at 4%, while the fifth expert estimates it at 14%. What should the committee's decision be? First scenario: the experts vote with a majority rule on the proposition "The risk of the additive is lower than 5% and can thus be authorized." By four against one, the proposition is accepted, and the additive is approved. Second scenario: the experts, who have perhaps read Lehrer and Wagner (1981), first take an equally weighted average of all experts' assessments, $(4 \times 4 + 14)/5 = 6\%$. As the result is higher than the admissible level, they decide to refuse the authorization. Is this a better decision?

This article is restricted to cases in which an expert panel is asked to make a decision on one binary question, which involves a comparison between a risk assessment and a risk threshold. This comparison may be explicit, for instance, there may be a given threshold as in the above food additive example, or implicit, as in the question "Is the risk-benefit ratio of this medicine worth it to be authorized for commercial use?" This is the actual sort of question posed to FDA advisory committees (Urfalino and Costa 2015). My aim is to assess which voting rule or aggregating function is the best one in these cases. This will, of course, require discussing what it could possibly mean to say that a voting rule is the "best."

The typical situation I consider is one in which experts take a vote (or engage in some aggregation procedure) after they have shared their views and deliberated. I am not considering the first step of information sharing and the way in which opinions change over the course of that process, but only the final aggregation step.[2] Hence, I shall assume that experts' beliefs do not change anymore and that no expert has any private information that has not already been shared with others. The latter assumption seems realistic insofar as expert committees are typically not just convened for one minute to vote on a decision, but are expected to discuss and share their views and perhaps converge on some points.[3]

1. The EFSA, the ECHA, and the European Commission actually first encourage panels to reach consensus, and majority voting is used if a consensus cannot be reached.

2. The distinction between these two steps is an idealization. In some cases, these steps are intertwined, as when the group reaches a decision by consensus or through the rule of nonopposition (cf. Urfalino 2014; Beatty 2017).

3. I briefly discuss the relaxation of this assumption in sec. 3.1. The assumption that quantities are well defined is discussed in sec. 3.2.

To study this problem more precisely, I shall model it as follows. The yes/no (or true/false) answer is supposed to be decided by comparing a risk assessment a to a risk acceptability threshold t, typically yes if and only if $a < t$. Both quantities are supposed to be quantitatively well defined at the individual agent level, for simplicity in [0, 1].[4] Typically, scientific expert panels are called for when there is no scientific consensus on a question matter, so in general, the n experts disagree on their individual risk assessments a_k ($k = 1, \dots, n$; n is supposed to be odd for simplicity). In sections 2–4, I adopt the assumption that experts have the same threshold t, either because they agree on it or because the institution imposes it. Note that this problem is formally the same as the one in which experts disagree on t and agree on a, and the results in these sections can be extended to this reinterpretation of the model. In section 5, I consider the more general scenario in which experts can have different risk assessments and different risk thresholds. Within this model, the aim of the article can be reexpressed as determining how the individual a_k's and t_k's should be best aggregated and compared, so as to give the group's binary answer to the question.

Note that although the model is presented in terms of a comparison between a "risk assessment" and "risk threshold," it can receive a much wider interpretation. Any question that involves a comparison between any two quantities, like costs versus benefits, can be modeled in that way. More generally, any binary question can be reconstructed as requiring an individual agent to assess whether answering "yes" is preferable to answering "no," that is, whether the utility of one answer is above the other: formally, the classical representation theorems of decision theory guarantee the existence of a utility function that applies to the two answers. So even if the binary question posed to a committee is not in the form of "Is a below t?" but of "Does X?", it can be reconstructed in this way and the present model applies.

The problem considered here is quite simple: the group has to answer only one question, which is binary, and individual beliefs are just over two variables (one of which may be fixed). However, it is already of much interest as it corresponds to many real-life cases: some expert panels are constituted on the sole purpose of answering one specific question or are asked to answer several but logically unrelated questions, for example, decisions about different medicines. Despite its simplicity, the problem is theoretically novel in several ways. First, it is different from the problems considered in

4. Traditionally, a risk assessment is the product of a severity and a likelihood. If both quantities are given quantitative assessments, then the risk is a nonnegative real number. For convenience, it can be mapped onto the interval [0, 1], for instance, with the function $x \to 1 - 1/(1 + x)$. So the results from this article can easily be generalized to a or t being real numbers. On the definition of the scale and reparameterization, see the discussion in sec. 3.

judgment aggregation theory, which studies the aggregation of binary judgments on an interconnected set of propositions (e.g., premises and conclusions) and has shown famous impossibility theorems, which imply, for instance, that the majority rule lacks certain desirable properties (for reviews, see List [2012] and Martini and Sprenger [2017]). Here, in contrast, an agent does not just have true/false beliefs on the matter of interest, but probabilistic ones (t_k and a_k), and the set of binary propositions is reduced to one element (the answer to the question). Second, our problem also differs from probability aggregation theory, which studies the aggregation of a set of probabilistic opinions (for reviews, see Dietrich and List [2016] and Martini and Sprenger [2017, sec. 3]). Since a threshold comparison is introduced, agents (and the group) do not hold just probabilistic opinions but also binary ones. Also, the set of probabilistic beliefs is quite limited here compared to what probability aggregation theory generally considers, and this proves to be of some importance, as discussed in section 2. Third, the model does not just amount to a threshold rule, according to which a binary belief is attributed to an agent just in case her degree of belief is above a threshold, since a group aggregation perspective is added. Finally, it is different from recent work by Dietrich and List (2018), which assumes an agenda larger than just one true/false proposition, as I do here.[5] The present article can be considered as a tentative bridge between these various frameworks.

The best decision rule for our binary question is likely to depend on the characteristics of the question, the experts, and the available knowledge and on other details. My methodological approach is not to conduct a detailed study of particular cases, but to look at features that most (interesting) cases share, so as to find general properties of the best decision rule.

The article is structured as follows. In section 2, I first argue that the framework of probability aggregation cannot help us solve the present problem. For the aggregation of scientific risk assessment on a specific question, a theory of its own is needed, and I try to sketch one here. In section 3, I argue that robustness considerations legitimate majority voting on the final decision, when one variable (a or t) is consensual. I then turn to the limits of majority voting. In section 4, I consider the case of independent risk factors. In section 5, I relax the assumption that experts agree on one variable and argue for the comparison of the medians. I conclude with some suggestions for improvements of scientific expert panels.

2. Probability Aggregation and Beyond. An individual expert can justify his or her yes/no answer to the binary question by providing his or her risk assessment a_k and comparing it to the threshold t (recall that, until sec. 4, t is assumed to be the same for all experts). If the group directly votes on the

5. An agenda is the set of propositions or items on which a judgment is to be made.

binary question, there is at first sight no group risk assessment that can be compared to t so as to justify its decision. Should one demand that such a group risk assessment $\tilde{a}$ can be reconstructed? Yes, for two reasons. First, normatively speaking, a group's opinion on the question is stronger (toward decision makers, the public, or other expert groups) if it can be justified by such a risk assessment than if it cannot. Otherwise, the group could be saying something like "We experts answer 'no' to the question of whether the risk is below the threshold, but it's not because we think it's above. Actually, we just don't have an opinion on what the risk is, and yet we do answer the question." The second reason is descriptive: currently, it is standard practice for agencies to require their expert groups to justify their views, not just to recommend a mere "yes" or "no." For instance, the rules of procedure of the Committee for Risk Assessment of the European Chemicals Agency mention that "the opinion adopted by the Committee shall consist of the position of the simple majority of all members present and having the right to vote, *including their grounds*" (ECHA 2015, art. 19-4; my emphasis). The group risk assessment $\tilde{a}$ in our model can be considered as a first step in this direction (a second step is taken in sec. 4 with the study of reasons). Also, when the FDA asks its panels questions like "Is the risk-benefit ratio of this medicine worth it to be authorized for commercial use?" the answer (for instance, "Yes, the risk-benefit ratio of this medicine is worth it to be authorized for commercial use") includes a mention of the risk-benefit ratio, on which the group should have a stand. For these reasons, I shall assume hereafter that it is a requirement for the group to be able to justify its binary decision with a group risk assessment $\tilde{a}$ to be compared with t. So that $\tilde{a}$ is not ad hoc, there should exist a general aggregation rule F that maps $\{a_1, \dots, a_n\}$ to $\tilde{a}$ (mathematically, F is a function from $[0, 1]^n$ to $[0, 1]$). In other words, our problem (when t is consensual) amounts to finding this aggregation function F. Note that the requirement is not that the group should state a group assessment $\tilde{a}$ (and then derive its binary decision from a comparison with t), but that the decision rule should allow for such an $\tilde{a}$ to be reconstructed.

Consider the majority rule. Is it consistent with the requirement that the group has a risk assessment $\tilde{a}$? In other words, if a group were to apply the majority rule on a yes/no vote, could an $\tilde{a}$ be reconstructed? Yes, for the following reason. The result of the majority vote is yes if and only if a majority of experts vote yes, that is, if and only if a majority of experts have a numerical assessment below the threshold, that is, if and only if the median of the experts' assessments is below the threshold. In other words, the majority voting rule on the decision is equivalent to considering the group's assessment $\tilde{a}$ as the median of the individual assessments a_k.

Besides the median, what are the other contenders for defining $\tilde{a}$? A standard way to aggregate numerical variables is by averaging. The linear aver-

age is defined as $\Sigma_k a_k/n$, and it can be generalized as $\Sigma_k \omega_k a_k$ with positive weights ω_k that sum to one, for instance, to take into account unequal degrees of expertise on the question.[6] Other averages are the geometric average or the harmonic average. As the example from the introduction has shown, these various probability aggregation rules can give different binary decisions for the group. Our goal is to determine which probability aggregation rule, followed by the threshold comparison, is the best one for our problem.

How to aggregate $\{a_1, \ldots, a_n\}$ into $\tilde{a}$ can be viewed as a probability aggregation problem, since the a_k lie in [0, 1]. Pooling probability functions has been studied for some time in the theory of probability aggregation (for reviews, see Dietrich and List [2016] and Martini and Sprenger [2017, sec. 3]). Can the results of this theory be used to select the best aggregation rule in our problem? Unfortunately the answer is 'no'. The framework of probability aggregation adopts an axiomatic approach: it starts by stating several axioms that seem to be desirable properties for the pooling function and then studies which function or aggregation rule, if any, satisfies them. The axioms considered in Dietrich and List's survey can be expressed in our case as follows:

> **Independence.** The group's probability a should depend only on the individual probabilities a_k.
>
> **Unanimity Preservation.** If all experts' probabilities a_k are the same, then the group's probability a should be this one too.
>
> **External Bayesianity.** If probabilities are to be updated on the basis of some information, it should make no difference whether the update occurs before aggregation (as all agents learn the information) or after aggregation (as the group learns the information).
>
> **Individualwise Bayesianity.** If probabilities are to be updated on the basis of some information, it should make no difference whether the information is received by a single individual before opinions are aggregated or by the group as a whole afterward.

6. This is akin to the iterated Lehrer-Wagner model (Lehrer and Wagner 1981, chap. 2), which, based on weights corresponding to the amount of respect agents have for one another, provides a single probability for the group. However, the iterated Lehrer-Wagner model, and especially its normative interpretation, have been subjected to many criticisms (for a survey, see, e.g., Martini and Sprenger [2017, sec. 4]). As a descriptive model, it is not useful for the present discussion. On assigning unequal weights to experts, see Klein and Sprenger (2015).

The Independence axiom is automatically satisfied here, because our problem does not contain any other probability on which a could depend. The Unanimity axiom is not desirable in cases in which agents have independent evidence for their beliefs. However, it is assumed that agents do not have any private information anymore, because they have extensively deliberated (see the introduction). Hence, the Unanimity axiom is a sound requirement here.

The axiom of External Bayesianity is relevant when some new item of evidence is acquired by the expert panel. In practice, the requirement makes sense only if the panel has remained the same when it has to deliver a second verdict, that is, if it is composed of the same individuals with the same individual probabilities. Whether it is the case that the same expert panel is asked to answer the same question twice is an empirical question.[7] This seems to be rarely the case: for instance, the FDA advisory committees generally include some temporary members and are renewed on a continuous basis. So the External Bayesianity axiom is not relevant in the cases we are concerned with.[8] The axiom of Individualwise Bayesianity makes sense if some new information is acquired by one of the agents, either externally or from another agent in the panel. The former case can be dismissed for the reasons just discussed, and the latter case is ruled out by the assumption that there is no private information anymore in the panel.[9] Note finally that the two Bayesian axioms do not make sense in the other interpretation of the model (see the introduction) where there is a consensus on a but not on the t_k. To sum up, the axiom of Independence is always satisfied, the Bayesian ones are not sound in our case, and the axiom of Unanimity is the only one that puts some constraint on the aggregation function.

A very large number of aggregation rules satisfy the Unanimity axiom: the median, linear averaging, geometric averaging, and so on—any convex function of the a_k. This illustrates the fact that a classical uniqueness result from the probability aggregation literature does not hold anymore: the well-known theorem by McConway (1981) and Wagner (1982), which states

7. I am not claiming that the normative force of an axiom comes only from its actual instantiations. It does make sense to require that some property should hold were some situations to occur. But these situations should be possible, and reasonably frequently so.

8. If it were relevant, it would speak in favor of geometric averaging and against the median (cf. Dietrich and List 2016, sec. 6). To anticipate: the robustness axioms I advocate in sec. 3 give opposite conclusions; hence, no aggregation rule would satisfy all axioms. One should then decide which of these axioms should be most important in the case under consideration.

9. If there were some private information, Individualwise Bayesianity would be a relevant axiom, and it would warrant multiplicative pooling (cf. Dietrich and List 2016, sec. 8). Here also, this axiom should be weighted against the robustness axioms of sec. 3.

that linear averaging functions are the only independent and unanimity-preserving functions. This is no surprise: the theorem requires a set of at least three events, whereas our problem considers only two: the product is risky, with probability a_k, and the product is not risky, with probability $1 - a_k$. The theory of probability aggregation has nothing to say on the aggregation of a single variable, and its results are useless for our problem. Considering a simpler agenda has widened the set of suitable aggregation rules, and no existing result can be used to pick the best one.

So how scientific expert panels should aggregate risk assessments is not a simple problem that can be solved straightforwardly with the existing literature, which has focused on general problems with complex agendas and has not addressed simple yet important questions. In the next section, I discuss other axioms or requirements that one may want to require so as to narrow the spectrum of possible aggregation rules.

3. Robustness Matters. Scientific expertise is supposed to meet some standards of truth or objectivity. This may be less meaningful or less important in political or daily contexts, to which judgment aggregation has been applied, but it is clearly a legitimate requirement in scientific contexts. For instance, as experts may make some mistakes in their individual risk assessments, it would be valuable that the aggregation rule diminishes, and does not exacerbate, the influence of these mistakes on the group's final decision. The rule should be sensitive to the core features of the problem, and not to tangential ones; that is, it should be robust to some changes that are regarded as irrelevant. This leads me to consider three elements with respect to which a decision could be robust: the risk metric, the level of detail, and the presence of strategic agents.

In the forthcoming discussion, some methods of averaging (linear, geometric, harmonic, hyperbolic) behave similarly, while the median (and its generalization with quartiles) behaves in a distinct way. To simplify the discussions, I will thus compare only the linear average and the median. The term $\mathcal{R}_m$ denotes the aggregation rule that compares the risk threshold with the median of the individual risk assessments (it is equivalent to a vote with the majority rule on the binary decision itself, or a supramajority rule for quartiles), and $\mathcal{R}_a$ denotes the aggregation rule that compares the threshold with the linear average (it also stands for the other averages).

3.1. Metric. The formal model I have introduced relies on a quantitative scale: a and t are given numerical values in [0, 1]. How is this scale defined in actual cases? My talking about probabilities has been only a matter of convenience (see n. 4), yet typical cases do not involve well-defined probabilities or explicit scales. For instance, a standard question posed to an FDA advisory committee is "Does the overall risk versus benefit profile for X sup-

port marketing in the US?" (Urfalino and Costa 2015, 183).[10] The formulation of this question lets experts identify what the appropriate risk versus benefit scale (or risk scale) is, and there is of course no unique way to do so.

There is a straightforward mathematical reason why there is no unique quantitative risk scale: any scale can be reparameterized by applying some transformation within [0, 1], such as $x \mapsto x^y$ with $y > 0$. One may also consider scales that are incommensurable with one another (Anderson 1993; Ackerman and Heinzerling 2004). Formally, an aggregation rule will be said to be robust against the scale metric if it is invariant under any positive monotonic transformation (i.e., a transformation that preserves the order).

This requirement poses problems for the rule $\mathcal{R}_a$. First, if there is individual heterogeneity about the risk scale, is it even possible for a chair to ask her colleagues "Please tell me your overall risk versus benefit assessment" given that each expert may have her own scale? Since the rule $\mathcal{R}_m$ is equivalent to majority voting on the binary question, it can be implemented without relying on individual numerical values and is thus safe from this criticism. Second, even if a common scale has been adopted, some theoretical difficulties remain. The outcome of a particular aggregation rule might depend on the scale that is employed, as shown in table 1. This dependence is a problem: which common scale should be chosen? (This is another aggregation problem!) Note that a variant of this problem exists even with a well-defined probability scale. For instance, let A be the event that a certain hazard (e.g., carcinogenic substances in food) is responsible for more than 10 cases of cancer in 100,000 people over 1 year. The experts estimate the probability of A, $p(A)$. Consider now A', the event that the hazard is responsible for more than 10 cases of cancer in 100,000 people over 10 years (as a whole, whatever the distribution along the years). Call $p(A')$ its probability. If the cancer cases are independent along the years, then $p(A') = 1 - [1 - p(A)]^{10}$. Because the relation between $p(A)$ and $p(A')$ is not linear, taking the linear average of the experts' assessments on A and transforming it into an assessment on A' or taking the linear average of the experts' assessments on A' does not give the same result. Which event A or A' should be considered is not clear—for instance, there seems to be no scientific argument to choose between them—and similarly for the right risk group assessment. However, there are other cases in which a common scale can be unambiguously defined.

10. As noted in the introduction, answering this question can indeed be reconstructed as comparing the quantities a and t as in the present model. However complex the risk vs. benefit profile is, a single answer yes or no has to be selected: a represents the overall risk vs. benefit assessment (for instance, obtained from some projection of the multidimensional original profile), whose existence is guaranteed by representation theorems from decision theory, and t is the threshold required for marketing.

TABLE 1. An Example of Outcomes

	a_1	a_2	a_3	t	Average a	$\mathcal{R}_a$	Median a	$\mathcal{R}_m$
x scale	.01	.01	.1	.05	.04	Yes	.01	Yes
x^2 scale	.0001	.0001	.01	.0025	.0034	No	.0001	Yes

Note.—The rule $\mathcal{R}_a$ gives different answers depending on the scale. The rule $\mathcal{R}_m$ is insensitive to the scale used.

For instance, this is the case for a subjective probability of some major event (with no ambiguous time consideration), such as whether the average global temperature will rise by more than 2°C in the next century.

So unless (i) the use of a risk metric is consensual among experts (either because it is given by the institution or because experts agree on it) and (ii) this metric is well justified (e.g., not as in the case of A and A'), the aggregation rule should be insensitive to the metric used to assess the risk and to express the threshold. Note that it is not enough that the institution provides a risk metric, for instance $p(A)$; the metric has to be justified in some way, so that the expert panel's answer does not depend on the institution's arbitrary choice and so that an ill-intentioned institution cannot take advantage of it.[11] The rule $\mathcal{R}_a$, which employs a linear average, does not fulfill this metric robustness requirement, as shown in table 1. The rule $\mathcal{R}_m$, which relies on the median, gives the same result regardless of a change of metric and is robust in this respect.[12]

3.2. Level of Detail. So far, an agent's assessment has been assumed to be a point value, that is, a number in [0, 1]. It may take other, less precise, forms: an interval or a qualitative judgment. Consider for instance the Intergovernmental Panel on Climate Change (IPCC) Assessment Reports, which regularly formulate a synthesis of existing scientific knowledge on climate change issues. The reports use a standardized vocabulary to express uncertainties, with several scales: some are quantitative, others are qualitative, and the latter have the advantage of being easily understandable by nontechnical audiences. Some scales are put in an explicit correspondence, as illustrated in table 2.

Thus, one may be interested in extending an aggregation function F defined on point values to other scales, such as interval or qualitative scales.

11. The justification need not be based on epistemic grounds only, and a social consensus can count as an acceptable justification. For instance, to decide whether a medical drug can be kept on the market, an institution may ask experts to assess the total number of people who have been severely injured by it so far, and not people who have died of it, if this is what society cares about.

12. The comparability of scales is also discussed in Risse's (2004) political philosophy work, and it is taken as an argument for majority voting.

TABLE 2. An Example of Correspondence

Term	Likelihood of the Outcome
Virtually certain	99%–100% probability
Very likely	90%–100% probability
Likely	66%–100% probability
About as likely as not	33%–66% probability
Unlikely	0%–33% probability
Very unlikely	0%–10% probability
Exceptionally unlikely	0%–1% probability

Note.—The scales are those used in the Fifth Assessment Report of the IPCC (Stocker et al. 2013, 142).

Formally, an ordinal scale S is given by a set of values, or items, and an order relation on it. Let $\mathcal{S}$ be the set of scales (including point values) to which the function F has been extended, and let F_S be the extension of F on a scale $S \in \mathcal{S}$, which maps $\mathcal{S}^n$ to $\mathcal{S}$. Assume that these scales are in correspondence: formally, S and S' are in correspondence (S being more precise) if and only if, for all $s \in S$, there exists $s' \in S'$ such that $s \subseteq s'$.

Which extensions of a given function F should be considered suitable for our problem? A plausible desideratum is that it should be robust against the level of detail of the scale that is used. If the individual assessments get more precise, so should the group assessment. That is, a group assessment obtained from more precise individual assessments should be compatible with that obtained from less precise individual assessments (if the precise group assessment was incompatible with the rough group assessment, the latter would not be meaningful). Technically, if the experts' assessments stand in an inclusion relation (e.g., 0.17 is included in [0.1, 0.2], itself included in the "unlikely" range in table 2), then so should the aggregated assessments. The function F should preserve inclusions:

Definition 1. An extended aggregation function F is *robust with respect to level of detail* just in case for all $S, S' \in \mathcal{S}^2$, for all $a_1, \dots, a_n \in \mathcal{S}^n$, for all $a'_1, \dots, a'_n \in \mathcal{S}'^n$,

$$\text{if } \forall\, k, a_k \subseteq a'_k, \text{ then } F_S(a_1, \dots, a_n) \subseteq F_{S'}(a'_1, \dots, a'_n). \tag{1}$$

Are our two candidates, the median and the linear average, robust with respect to level of detail? To avoid some definition problems, let us assume that within a scale, items do not overlap (for all $S \in \mathcal{S}$, for all $x, y \in S^2$, $x \cap y = \varnothing$). Extending the median to any kind of scale is straightforward, as one can show (proof in the appendix):

Proposition 1. The median is robust with respect to level of detail.

There are several ways to extend the linear average to other scales. For instance, in table 2, which number should be associated with the 90–100 interval when averaging? The metrics considerations from the previous section apply here forcefully too and warn us against the possible arbitrariness of a straight "95" average. To avoid this, the extension of the linear average to other scales should be both consensual among experts and well justified (in a similar way to the previous section). Assume this is the case, and consider one plausible candidate for an interval scale: the linear average of intervals is defined as the interval that includes the linear average of the middles of the intervals. Call this extension ϕ. One can show (proof in the appendix):

> **Proposition 2.** The linear average extended to ϕ is not robust in general against the level of detail. It is robust just in case all intervals of the scale are of equal size.

This result might be seen from a glass-half-empty or a glass-half-full perspective. However, the latter relies on the notion of "equal size," and for it to be nonarbitrary, the previous section tells us that the metric has to be consensual and well justified.

Overall, linear averaging is not robust in general for the level of detail; it is in the limited case in which the metric of the numerical scale and the choice of the extension (for instance, the middle of intervals) are all both consensual and well justified. In practice, this leads to a lot of conditions: these are all prerequisite aggregation problems! On the other hand, $\mathcal{R}_m$ is robust in general for the level of detail.

3.3. Bias and Strategic Votes. Not all experts are moved by epistemic goals only, and conflicts of interest can arise. For instance, numerous controversies have surrounded FDA advisory committees along the years (Urfalino and Costa 2015, 168–69). If a better selection of experts may be part of the solution, which decision rule is used in the expert panel can also play an important role in reducing, or enabling, the impact of biased agents. For instance, with $\mathcal{R}_a$, an expert who is paid by a pharmaceutical firm can strategically express a much lower risk of a medicine to influence the group's average: with a threshold at 10%, she might express 0.1% instead of the 9% that actually reflects her sincere belief. That is, the rule should be such that no agent can achieve a collective binary decision that is closer to her individual binary opinion by misrepresenting rather than truthfully revealing her risk assessment.[13] This requirement is all the more important given that a biased ex-

13. For more precise definitions of this, see Balinski and Laraki's (2010, 189–90) "strategy-proofness in grading" or List and Pettit's (2011, chap. 5) "incentive-compatibility." In our simple case, any conception is acceptable.

pert may have already influenced the panel during the discussion preceding the vote. A rule that satisfies this robustness requirement is also more objective according to two senses in Douglas's (2004) classification: (i) detached objectivity, which says that one's personal values (e.g., allegiance to a firm) should not prevail on evidence (e.g., that the probability is indeed 9%, as above); and (ii) procedural objectivity, according to which the same result obtains, no matter who is involved in the process, whether strategic or not.

As the above example shows, the rule $\mathcal{R}_a$ is not robust against strategic agents: an expert who wants the group to make some specific decision can give a higher or lower assessment. This is possible because the average takes into account the distance at which experts' assessments lie. This is not the case for the majority rule, since there is no way to vote more or less yes: strategy-proofness is indeed a well-known feature of the majority rule, and here of $\mathcal{R}_m$.

3.4. Conclusion. I have considered three ways in which the decision made by a scientific expert panel could be more objective and robust: against the metric used for the risk scale, the level of detail, and strategic voting. Although I have argued that these are sound requirements in general, one may not always require all of them in every case (for instance, I have discussed in which cases the metric robustness is not needed). A clear result stands out: each of these robustness requirements favors $\mathcal{R}_m$ over $\mathcal{R}_a$. Recall that the former compares the threshold with the median of experts' assessments and is equivalent to a vote with the majority rule on the binary question directly, while the latter compares the threshold with the average. This robustness study provides a substantial justification for the traditional majority rule that expert panels often use when confronted with a binary decision.

Compared to probability aggregation and section 2, an important difference should be noted: while linear averaging is justified on solid grounds for a large agenda, it becomes undesirable (worse than the median) as soon as the agenda is narrowed to two options only. This shows that precisely delimiting the agenda on which experts are asked to give an opinion is of crucial theoretical matter. It should be an important step when one applies aggregation theory to real cases. Also, the present results suggest that it would be interesting to consider the above robustness requirements in cases of large agendas; impossibility theorems could be expected because of the conflict between the median, which is more robust, and the average, which is supported by standard axioms.

4. Reasons. So far, I have argued that a minimal requirement is that a group risk assessment $\tilde{a}$ can be reconstructed to justify its yes/no decision by comparing it to the threshold. In some situations, one may require that the group provides a deeper justification. In this section, I extend the previous model

by considering some reasons for the justification. I shall assume (cf. sec. 3.1) that the risk scale is consensual among experts and well justified. Suppose experts agree that the risk under consideration is determined by m independent factors ($m \geq 2$), of probability a_j ($j = 1, \ldots, m$)—for instance, the risk associated with a medicine comes from m unrelated secondary effects. Then a is the probability that at least one risk factor triggers:

$$a = 1 - \prod_{j=1}^{m}(1 - a_j). \tag{2}$$

Suppose each expert k has her own assessment of each factor $a_{k,j}$ ($j = 1, \ldots, m$). The problem is then to aggregate the $n \times m$ matrix of probabilities $a_{k,j}$ and to compare the result with the threshold t.

As the m factors are independent, a sound requirement is to aggregate the individual assessments of them separately. How should that be done? Adapting the robustness arguments from the previous section (except sec. 3.1) leads to the conclusion that the panel should take the median of the individual assessments for each factor. However, there is a fundamental limitation to this: as $m \geq 2$, the conditions of the theorem by McConway and Wagner (cf. sec. 2) are now fulfilled, so its conclusion applies:[14] the only acceptable probability aggregation rule on the set of factors and on the overall decision is linear averaging. In other words, if the group uses the median to determine both the independence factors' values and the overall risk (according to the above results), then this does not give a probability function and inconsistencies can arise. Table 3 gives such an example. Asking the expert panel to take stands on the reasons for its majority decision can lead it to change its decision.

Does this mean that our robustness defense of the median should be discarded? Not necessarily. The theorem by McConway and Wagner assumes that the experts aggregate their views both on the independent factors and on the overall risk assessment. But one can have the experts aggregate their views on the independent factors only.[15] The overall risk assessment is then computed according to equation (2), and the final decision is obtained by comparing this value to the threshold. In that way, experts do not vote on the final decision directly. This decision rule is a so-called premise-based

14. Each of the $m \geq 2$ factors can be triggered or not, so there are at least four events, which is higher than the three required in the theorem. Also, requiring that the aggregation proceeds on each factor independently is just requiring the classical Independence axiom. The axiom of Unanimity Preservation is also assumed here.

15. In doing so, the axiom of Independence should be replaced by the weaker axiom of "Independence on premises" (Dietrich and List 2017).

TABLE 3. A Case in Which the Rule of the Median Can Lead to Inconsistencies

Risk Aspect	a_1	a_2	$a = 1 - (1 - a_1) \cdot (1 - a_2)$
Agent 1	.01	.01	.0199
Agent 2	.02	.01	.0298
Agent 3	.01	.02	.0298
Median	.01	.01	.0199 or .0298?

Note.—With a threshold at 0.025, the group's decision could be either yes or no.

rule.[16] Then the linearity result of McConway and Wagner does not apply any more. The robustness considerations from the previous section do apply at the level of independent factors, and they recommend that the group takes the median of the individual assessments. Afterward, the total risk should be computed from these aggregates using equation (2) and compared to the threshold.

5. When Experts Do Not Agree on the Risk Threshold Either. Suppose now more generally that both risk assessments a_k and thresholds t_k may vary among experts. How should experts aggregate their views and reach a collective decision, yes or no? Here are some possibilities: simply apply the majority rule on the experts' binary votes or collect the individual a_k and t_k and declare that the group's decision is yes just in case the linear average of the a_k is below that of the t_k. Or similarly with the median. Which aggregation rule is the best one (according to which criteria)?

The case of the majority rule requires special attention. In section 2, when t is consensual, it has been noted that the majority rule on the binary question is equivalent to the comparison of the median of $\{a_1, \ldots, a_n\}$ with t. When individuals disagree on t, is the majority rule equivalent to a comparison of the median of $\{a_1, \ldots, a_n\}$ with the median of $\{t_1, \ldots, t_n\}$? Proposition 3 answers negatively (proof in the appendix).

> **Proposition 3.** There is no implication relation in one sense or another between
>
> i) a majority vote on whether $a < t$ and
> ii) the median of $\{a_1, \ldots, a_n\}$ being smaller than the median of $\{t_1, \ldots, t_n\}$.

Thus, when neither a nor t is consensual among experts, the median loses its privileged status with respect to majority voting. It makes a difference

16. On this strategy, see Cooke (1991), Bovens and Rabinowicz (2006), Hartmann and Sprenger (2012), Bradley, Dietrich, and List (2014), and Dietrich and List (2017).

whether experts compare-then-aggregate (with majority voting) or aggregate-then-compare (with the median). In other words, when a group reaches a binary opinion through majority voting on the proposition, it is not possible to reconstruct a group's view $\tilde{a}$ and $\tilde{t}$ as the median of the individual assessments.

Can this be done through another function F than the median? Recall the importance of this feature from the beginning of section 2. The function should satisfy some basic desiderata, at least the traditional anonymity requirement (in line with what is traditionally required in judgment aggregation, see List [2012]):

> **Definition 2.** An aggregation function F from $[0, 1]^n$ to $[0, 1]$ is anonymous if, for any profile $(x_1, \dots, x_n)$, $F(x_1, \dots, x_n)$ is invariant by any permutation of the profile.

If a function F is anonymous, who holds which view x_k does not matter; only which views are held does. This is a sound requirement in expert panels when all experts have equal weight. However, the following result can be shown (proof in the appendix):

> **Proposition 4.** There is no anonymous aggregation function F from $[0, 1]^n$ to $[0, 1]$ such that a majority vote on whether $a < t$ is equivalent to $F(a_1, \dots, a_n) < F(t_1, \dots, t_n)$.

In other words, the result of a majority vote on whether $a < t$ cannot be reconstructed in general through $\tilde{a}$ and $\tilde{t}$ that are the result of the same anonymous aggregation function—although every individual has an opinion on a and t. So using the majority rule requires the group to abandon the most basic requirement of being able to justify its decision as a group by saying that the risk is indeed assessed to be below the threshold. This makes majority voting lose much appeal in scientific expert committees as soon as the two variables a and t are not consensual.

But it gets worse. The majority rule does not respect another desirable requirement related to objectivity considerations. Once the scientific experts have been selected by the institution, only arguments should be taken into account, and who holds which views should not matter for the group's decision. This should also hold in detail; that is, the decision rule should be invariant with respect to permutations on which individuals hold which view on either a or t. If the two variables a and t were logically or mathematically related, it would not make sense to ask for permutations of partial views among individuals. But in the case of a committee deciding on a risky question, no particular relation between a and t should exist. The risk assessment an expert makes, which is a factual judgment, should not be directly influ-

enced by the risk acceptability threshold she has, which involves a value judgment. This is what Douglas (2004) expresses as the requirement of detached objectivity. Hence, the following variant of the anonymity axiom is also desirable:

> **Definition 3.** An aggregation function G from $[0, 1]^{2n}$ to {yes, no} is detailed-anonymous if for any two profiles $(a_1, \dots, a_n)$ and $(t_1, \dots, t_n)$, $G(a_1, \dots, a_n, t_1, \dots, t_n)$ is invariant by any permutation on either $(a_1, \dots, a_n)$ or $(t_1, \dots, t_n)$.[17]

Requiring detailed-anonymity means that what matters is the set of individual views in the expert panel on a and on t, not how these individual views are spread over the experts and whether some are held by the same expert. Suppose that instead of gathering n experts who have opinions on both a and t, the institution gathers $2n$ experts, half of which have an opinion on a and half of which have an opinion on t. The detailed-anonymity axiom requires that the group's opinion should be the same in both cases and thus should not depend on how these $2n$ experts are actually paired to yield n experts.

Then, one can show the following proposition (proof in the appendix):

> **Proposition 5.** There is no detailed-anonymous aggregation function G from $[0, 1]^{2n}$ to {yes, no} such that a majority vote on whether $a < t$ is equivalent to $G(a_1, \dots, a_n, t_1, \dots, t_n)$.[18]

In other words, the majority rule does not respect detailed-anonymity; more precisely, the majority rule applied on binary judgments cannot be reconstructed as an aggregation function that is detailed-anonymous on the risk and threshold assessments. This gives another reason to reject the majority rule when there is no consensus on either a or t in the expert panel. How should the expert panel aggregate the individual views of its members, if not with the majority rule? As I have argued that one should be able to reconstruct a group risk assessment and a risk acceptability threshold, the discussion of section 3 applies to each quantity: the median should be preferred over other kinds of averages. Then the group's binary decision should logically follow from a comparison between these medians. In other words, when experts disagree on the risk assessment and on the threshold, they

17. Note that this function G is more general than the function F considered in proposition 4, as it does not apply separately to either $\{a_n, \dots, a_n\}$ or $\{t_n, \dots, t_n\}$. For instance, it could require that each a_k is below all t_i.

18. Note that proposition 5 is more general than proposition 4: the latter can be derived from the former with $G =$ truth value$\{F(a_1, \dots, a_n) < F(t_1, \dots, t_n)\}$.

should compare the medians of their individual assessments. This obviously respects anonymity and detailed anonymity and enables the group to have meaningful $\tilde{a}$ and $\tilde{t}$.[19]

6. Conclusion. This article has investigated the rationale for the majority rule that is often used in scientific expert panels. To this end, I have introduced a model of individual decisions, which relies on a comparison between a risk assessment and a risk acceptability threshold. Four main points have been shown.

1. The standard framework of probability aggregation cannot solve the problem.
2. When the threshold is common to all experts, robustness considerations clearly favor majority voting on the decision, which is equivalent to comparing the threshold with the median of the individual risk assessments.
3. When the risk assessment comes from several independent factors, the median rule can conflict with a majority vote on the final binary decision. The group should then aggregate each factor with the median.
4. When experts disagree on both the risk assessment and the acceptability threshold, majority voting on the binary decision does not respect, in general, some basic requirements: enabling the group to have an opinion on both a and t and being insensitive to the details of which expert holds which view. Instead, the experts should compare the medians of their assessments of a and t.

When making a binary choice, the majority rule often goes unquestioned, perhaps because of an overly simplistic modeling of the problem in which agents have a binary belief (or preference). The present results show that adopting a slightly richer and more realistic model with a threshold comparison implies that the majority rule is not the best aggregation rule. In sum, the

19. When the scale is ordinal and contains a few grades, the comparison of the medians resembles Balinski and Laraki's (2010) "majority judgment." Let us note an important difference, though. The interpretation of the rating scale is different: for Balinski and Laraki, a grade is attributed to a candidate and, more generally, to a decision that could be made (e.g., choose X for president, authorize this drug or don't). Here, a grade is not attributed to a possible decision, but to some quantity that has no direct link with it, namely, a risk assessment or an acceptability threshold. In other words, a and t cannot be interpreted as grades of the possible decisions, and so there is no direct formal similarity between Balinski and Laraki's model and mine. Comparing the two would require another paper. In an interesting paper, Morreau (2016) discusses limitations to using grade assignments in groups due to threshold uncertainty, in relation with Balinski and Laraki's proposal, and is thus somehow between their framework and mine.

defense of the majority rule is not robust with respect to the further theoretical considerations that were added to the model.

These theoretical results suggest some concrete improvements for scientific expert panels. First, as the robustness discussion in section 3.1 has shown, an interesting possibility for the institution that gathers an expert panel is to explicitly offer a well-justified metric. Second, if the institution suspects that no consensus will emerge on either the risk assessment or the threshold, the institution should not require the use of the majority rule but the comparison of the medians. To use the majority rule, a sufficient condition is that the institution provides the threshold to the committee. As the threshold is a value judgment, a democratic institution might (roughly) take it from society and give it to the experts rather than decide on it alone. Third, the institution should encourage expert panels to divide questions and aggregate their estimates on causal factors first instead of making a final decision directly. When the FDA asks its advisory committees to take majority votes on questions of the form "Does the overall risk versus benefit profile for X support marketing in the US?", it could benefit from all these suggestions.

Appendix

Proof of the Propositions

Proof of Proposition 1. Reindex the a'_k's with a permutation function such that they are in increasing order. Assuming n is odd, $n = 2m + 1$, a'_{m+1} is the median of the a'_k's. Reindex the a_k's with the same permutation function. Because $a_k \subseteq a'_k$ and a'_k do not overlap, the a_k's are now also in increasing order. So a_{m+1} is the median of the a_k's, for the same m. Hence, $a_m \subseteq a'_m$ by hypothesis. QED

Proof of Proposition 2. Suppose the intervals of the scale are not all of equal sizes. Consider two contiguous intervals of different sizes, and suppose without loss of generality that the second one is larger, in the form $a'_1 = [x, x + \delta[$ and $a'_2 = [x + \delta, x + 2\delta + \epsilon[$, with δ and ϵ in $]0, 1]$. Suppose $a_1 = x$ and $a_2 = x + \delta$. Then the linear average is $F_S(a_1, a_2) = x + \delta/2$, which lies in a'_1. The linear average of the middles of the intervals is $x + \delta + \epsilon/4$, which lies in a'_2, not in a'_1. So the robustness condition is not fulfilled.

Suppose now that the intervals are of equal size s. Note that the difference between an individual assessment and the middle of the interval to which it belongs is bounded by $s/2$. So the difference between averages of these quantities is also bounded by $s/2$. Consequently, there is the inclusion to be demonstrated for the robustness condition.

Proof of Proposition 3. Proposition 3 is a special case of proposition 4 when the aggregation rule is the median, which is an anonymous rule.

Proof of Proposition 4. Suppose, for reductio, that an anonymous function F exists. By simplicity, suppose the number of experts is odd: $n = 2m + 1$. Consider the following case: the first $m + 1$ experts vote yes, that is, $a_1 < t_1, \dots, a_{m+1} < t_{m+1}$, while the m others vote no, that is $t_{m+2} < a_{m+2}, \dots, t_n < a_n$. Thus, a majority vote yields a yes. Assume also that $t_{m+2} < a_1 < t_1 < a_{m+2}$ (which is consistent with the above relations).

Suppose now a_1 and a_{m+2} are permuted (but not t_1 and t_{m+2}). Because F is anonymous, it gives the same group values for $\tilde{a}$ and $\tilde{t}$ with or without the permutation. For the first agent, $a_{m+2} > t_1$ (because of the assumption made above), so she votes no instead of yes. For agent $m + 2$, $a_1 > t_{m+2}$, so she still votes no. As one yes has turned into a no, the majority switches for no, which contradicts the fact that the same relation holds for $\tilde{a}$ and $\tilde{t}$.

Proof of Proposition 5. Suppose, for reductio, that there exists such a G function. An argument similar to the proof of proposition 4 can be made, which yields a contradiction.

REFERENCES

Ackerman, Frank, and Lisa Heinzerling 2004. *Priceless: On Knowing the Price of Everything and the Value of Nothing*. New York: New Press.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Balinski, Michel, and Rida Laraki. 2010. *Majority Judgment: Measuring, Ranking, and Electing*. Cambridge, MA: MIT Press.

Beatty, John. 2017. "Consensus: Sometimes It Doesn't Add Up." In *Landscapes of Collectivity*, ed. Snait Gissis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.

Bovens, Luc, and Wlodek Rabinowicz. 2006. "Democratic Answers to Complex Questions: An Epistemic Perspective." *Synthese* 150:131–53.

Bradley, Richard, Franz Dietrich, and Christian List. 2014. "Aggregating Causal Judgments." *Philosophy of Science* 81:491–515.

Cooke, Roger M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: Oxford University Press.

Dietrich, Franz, and Christian List. 2016. "Probabilistic Opinion Pooling." In *Oxford Handbook of Probability and Philosophy*, ed. A. Hájek and C. Hitchcock. Oxford: Oxford University Press.

———. 2017. "Probabilistic Opinion Pooling Generalized Part Two: The Premise-Based Approach." *Social Choice and Welfare* 48 (4): 787–814.

———. 2018. "From Degrees of Belief to Binary Beliefs: Lessons from Judgment-Aggregation Theory." *Journal of Philosophy* 115 (5): 225–70.

Douglas, Heather E. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138:453–73.

ECHA. 2015. "Rules of Procedure for the Committee for Risk Assessment." Management Board Decision 21/2015. Ref. MB/19/2015 final, Part 1 RAC. https://echa.europa.eu/documents/10162/13579/rac_rops_en.pdf.

Hartmann, Stephan, and Jan Sprenger. 2012. "Judgment Aggregation and the Problem of Tracking the Truth." *Synthese* 187:209–21.

Hauray, Boris, and Philippe Urfalino. 2007. “Expertise scientifique et intérêts nationaux: L’évaluation européenne des médicaments 1965–2000.” *Annales HSS* 2:273–98.
Klein, Dominik, and Jan Sprenger. 2015. “Modelling Individual Expertise in Group Judgments.” *Economics and Philosophy* 31:3–25.
Lehrer, Keith, and Carl Wagner. 1981. *Rational Consensus in Science and Society*. Dordrecht: Reidel.
List, Christian. 2012. “The Theory of Judgment Aggregation: An Introductory Review.” *Synthese* 187:179–207.
List, Christian, and Philip Pettit. 2011. *Group Agency*. Oxford: Oxford University Press.
Martini, Carlo, and Jan Sprenger. 2017. “Opinion Aggregation and Individual Expertise.” In *Scientific Collaboration and Collective Knowledge*, ed. T. Boyer-Kassem, C. Mayo-Wilson, and M. Weisberg. Oxford: Oxford University Press.
McConway, Kevin J. 1981. “Marginalization and Linear Opinion Pools.” *Journal of the American Statistical Association* 76 (374): 410–14.
Morreau, Michael. 2016. “Grading in Groups.” *Economics and Philosophy* 32:323–52.
Risse, Mathias. 2004. “Arguing for Majority Rule.” *Journal of Political Philosophy* 12 (1): 41–64.
Stocker, Thomas F., et al. 2013. *Climate Change 2013: The Physical Science Basis; Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
Urfalino, Philippe. 2014. “The Rule of Non-opposition: Opening Up Decision-Making by Consensus.” *Journal of Political Philosophy* 22 (3): 320–41.
Urfalino, Philippe, and Pascaline Costa. 2015. “Secret-Public Voting in FDA Advisory Committees.” In *Secrecy and Publicity in Votes and Debates*, ed. Jon Elster, 165–94. Cambridge: Cambridge University Press.
Wagner, Carl. 1982. “Allocation, Lehrer Models, and the Consensus of Probabilities.” *Theory and Decision* 14:207–20.