Philosophers' Imprint

VOLUME 25, NO. 14
JULY 2025

RESOLUTE AND CORRELATED BAYESIANS

Boris Babic, Anil Gaba, Ilia Tsetlin, Robert L. Winkler

University of Hong Kong,
INSEAD,
INSEAD,
Duke University, Fuqua School of Business

doi.org/10.3998/phimp.3416

1. Introduction

In this project we reframe the belief aggregation problem as an evidence combination problem. We explain that the focus on combining credences alone, as is the case in prevailing approaches, ignores the individual evidential states giving rise to those credences. As a result, traditional approaches fail to capture the multitude of individual evidential states which can lead to the same group credences. This occurs when we fail to account for dependence among individuals and the resilience of their beliefs. Such omissions are not innocuous: they can underdetermine both the group belief and its updating strategy.

We present an approach that allows one to focus instead on appropriately combining evidence, and in particular taking into account any overlaps in information. Once the evidence is properly captured, we will show, a full group distribution can be uniquely established on its basis. From this distribution, we can derive point estimates, intervals, and predictions. We call this the evidence-first method, in part to distinguish our approach from prevailing rules for combining beliefs, which may more accurately be described as credence-first.

To understand what we mean by this distinction – between combining evidence and combining credences – consider an example: Ahmed has observed 100 coin tosses, 30 of which were heads. His estimate that the next toss will land on heads is 0.3. Beatrice has observed 10 tosses, 3 of which were heads. Her estimate for heads on the next toss is also 0.3. Ahmed and Beatrice have identical probability estimates but very different information states, which we will characterize in terms of Joyce (2005)'s notion of resilience. This is not a distinction without a difference: which evidential state the probability is based on can have a profound effect on how the group responds to and acts on new information. In the above example, ordinary averaging suggests a group belief of 0.3, but it does not say whether this belief corresponds to 30/100 heads, 3/10 heads, or 33/110 heads. Each alternative would lead to very different updating behavior. Matters become even more complicated when some of the tosses that form the basis of Ahmed

and Beatrice's evidence were observed in common – i.e., when their information is overlapping and their estimates are correlated – a situation that is ubiquitous in real life (Lindley, 1983). Thus, when we seek to combine Ahmed and Beatrice's beliefs we need to know, first, the evidence they correspond to and, second, the extent of its overlap. The prevailing approaches in the literature are not equipped with the tools to answer these questions. The evidence-first method can answer them.

The paper proceeds as follows. In Section (2), we provide a brief background and motivate the general project. In Section (3), we reframe the belief aggregation problem. We explain that its usual formulation – combining a list of credence functions into a group credence function – is underspecified. Without capturing the size or weight of each member's evidence – i.e., the resilience of their credences – the group belief fails to take into account the full texture of information available to its members. In Section (4), we develop the evidence-first method. While this gets a little bit technical, the basic idea is very simple and intuitive: it is just a matter of properly accounting for everyone's evidence – do not leave anything out, do not overcount. Consider the Ahmed and Beatrice example above. If they did not observe any tosses in common, then the group's evidence consists of 110 tosses, 33 of which were heads. If they observed some tosses in common, we must appropriately subtract these. We then explain how the evidence, together with each person's prior, fixes a unique group distribution that captures the probabilities (valence), their resilience (weight or sharpness) and the dependence among individuals (correlation). In Section (5), we provide a fully worked example of our approach. And in Section (6), we explain how this approach can be generalized to provide normative guidance even in cases where we do not have full access to the individuals' underlying evidence.

2. Background

We are interested in how beliefs from multiple individuals ought to be combined to form a group belief. This problem can manifest in several different ways. The first and most literal is when we inquire into the opinions of a collective, taken as one agent (List and Pettit, 2011). This may occur when the collective is the subject of reactive attitudes like praise or blame (Strawson, 1962). For example, Amnesty International may blame Shell for human rights abuses in Nigeria without necessarily singling out a corrupt set of individuals to bear responsibility. A judge may decide that a corporation entered into a contract even if no particular set of individuals explicitly thus intended. Indeed, the legal notion of corporate personhood requires that we impute agency to corporate entities. As Chief Justice Marshall states in a well-known case before the US Supreme Court, "The great object of an incorporation is to bestow the character and properties of individuality on a collective and changing body."[1]

The second is when an individual needs to combine multiple sources of counsel or advice. For instance, Alibaba co-founder Lucy Peng is deliberating whether to purchase a small but promising venture. She solicits advice from three different domain experts on whether the company will turn a profit in five years. After obtaining their estimates, she must combine them into one prediction about profitability which represents her own credence. And third is when a group must act. For example, Tesla's board of directors must decide whether to remove its founder Elon Musk as the company's CEO. Before they can make a decision, they need to combine their individual beliefs about the wisdom of doing so.

The prevailing rules in the aggregation scholarship use measures of central tendency to identify a group's belief. Moss (2011) and Pettigrew (2019), for example, defend ordinary averaging whereas Russell et al. (2015) and Dietrich (2019) champion geometric averaging. Dietrich and List (2016) discuss the multiplicative rule, which is a special case of the latter. We will explain how measures of central tendency can arise naturally from considerations of evidential symmetry under our approach. However, depending on the underlying evidential states, our approach may or may not coincide with any form of averaging.

1. *Providence Bank v. Billings*, 29 U.S. 514 (1830).

Meanwhile, Easwaran et al. (2016) use the product of odds ratios, which is somewhat closer to our approach.[2] Indeed, we will explain that the rule they develop is equivalent to our method under special circumstances (independent signals and a uniform prior).[3]

While the prevailing aggregation methods in Bayesian epistemology largely focus on measures of central tendency, there are some views closer to ours which can be found in the logic of belief revision literature. For example, Williamson (2019) argues that the group distribution should be the distribution which maximizes Shannon information entropy subject to the constraints imposed by the evidence of each of the agents. Our approach is in the spirit of Williamson's, as we too start from the motivating idea that the content which should be combined is the agents' evidence.

However, Williamson does not explore situations of evidential overlap. In this project, those are the most interesting situations, and the ones that we spend the most time developing. When evidential bases are independent, aggregation is easier, and even the prevailing averaging rules yield intuitive results. It is particularly in cases of overlap where things get tricky, and our approach attempts to address them. In that sense, one can think about our project as constructively building on Williamson (2019)'s. However, we combine probability distributions in a different way – we do not use maximum entropy methods. In that sense, our project is doing something different, though in the same spirit.

When we consider cases of evidential overlap, the aggregation problem becomes particularly interesting. Our approach requires that the individuals in the group can share evidence with each other and determine which bits are overlapping and which bits are not. For example, Williamson considers a case where we have two doctors making a prognosis about a patient's cancer, where one doctor has clinical evidence and the other doctor has molecular evidence. This is a nice case for our project as well, but it is arguably an easy case. Here the doctors can share their evidence and there is no overlap. We can modify the example so that there is some shareable overlapping evidence though. For example, perhaps both doctors physically examined the patient (measuring their temperature, blood pressure, etc.). Both the original example and this modification are ideal use cases for our model, because here the overlapping evidence can be easily identified. However, imagine a case where a forecaster must combine two analysts' predictions, without knowledge of the underlying evidence that the predictions were based on. In this case, we cannot aggregate evidence since we do not know what it is or the extent to which it overlaps among the analysts' who made the predictions. We consider this situation in Section 6 of the paper, and we explain that even though in such cases our model cannot provide a recipe, so to speak, for combining credences, it can be used as a normative benchmark for which combinations are reasonable and which are not.

3. Group Credence is not Reducible to Valence

The argument in this section is straightforward. The prevailing combination rules – those which rely on one type of averaging or another – fail to capture an important aspect of the aggregation problem's information structure: namely, the weight or mass of the group members' credences, which we call their resilience and define more carefully below.

Let $X : \mathcal{F} \to \mathbb{R}$ be a random variable defined on an underlying σ-algebra $(\Omega, \mathcal{F}, P)$. Let $c(x)$ and $C(x)$ be the individual and group credence functions for X. Then ordinary and geometric averaging may be defined as follows.

2. Kinney (2020) moves away from averaging and uses stacking, which is a particular case of ensembles from machine learning. This is a different spirit of aggregation, but it would be hard to apply in cases where there is not much data or when, as often, the future is expected to be significantly different from the past – i.e., where so-called concept drift occurs (Widmer and Kubat, 1996).
3. Throughout this project, we use evidence, information, data, and signal interchangeably. In the mathematical portions, it will be unambiguous what the evidence is.

Ordinary Averaging:

$$C(x) = \sum_{i=1}^{n} w_i c_i(x).$$

Geometric Averaging:

$$C(x) = \prod_{i=1}^{n} c_i(x)^{w_i}.$$

Simple (ordinary/geometric) averaging is obtained from (ordinary/geometric) averaging by setting all weights $w_i = 1/n$. The so-called multiplicative mean is obtained from the geometric mean by setting $w_i = 1$. Some authors also suggest normalizing the result, so that the group credence is given by the above equations multiplied by their normalizing factor. All of these rules share the following important property:

> **Credence profile sufficiency**. An individual's list of probability assignments, which Dietrich (2019) calls their credence profile, is a sufficient statistic for summarizing their doxastic contribution to the group's belief.

For example, suppose we are interested in identifying a group's probability that the next ball to be drawn from a certain urn will be white.[4] The urn contains blue and white balls in unknown proportion. Credence profile sufficiency says that what we need from every individual is a list containing the probability that the next ball to be drawn is white, and the probability that the next ball to be drawn is blue. Or, equivalently, their point estimates of the proportion. For example, A's list for (White, Blue) might be (0.6, 0.4). We will argue that the credence profile is not enough for identifying the group belief because there can be many group credences that map back to the same credence profile depending on the underlying members' evidence. This becomes particularly clear when the group undergoes a learning experience, in which case the group update is often underdetermined as well.

Joyce (2005), following Skyrms (1980), distinguishes between the valence, on the one hand, and resilience (mass or weight), on the other, of a credence function. Valence, Joyce says, "is a matter of which way, and how decisively, the relevant data points" (p. 159). Meanwhile, the "size or weight of the evidence has to do with how much relevant information the data contains, irrespective of which way it points" (p. 159). We refer to the latter as its resilience. And we refer to people with more/less resilient credence functions as more/less resolute, for short.[5]

Just as a vector has a direction and a magnitude, so too does a credence have a valence and a resilience. The valence refers to its direction, as an estimate of a proposition's truth value or an event's likelihood of occurring. A credence of 0.9 that it will rain has a strong valence in favor of rain. By defining a group's credence as a function of the credence profile – i.e., the list of individual credence functions – the prevailing approaches essentially combine individual valences into a group valence. But in doing so they neglect weight or resilience. This is an especially glaring omission when *combining* credences where we want to pool every individual's full contribution, which may vary from person to person, depending on their level of expertise or background information.

4. We use ball-and-urn examples throughout. While these are not the most exciting, they are flexible, the evidence is unambiguous (i.e., observed balls) and they allow us to neatly describe various group learning scenarios. In Section (5), we use a more realistic example to illustrate our approach.

5. For Joyce, resilience is to be understood in terms of the extent to which a person's credences change under new data. But resilience is not a purely diachronic concept – we will explain that it can be captured from a time-slice centric perspective also, to borrow Moss (2015) and Hedden (2015)'s expression.

To characterize resilience, and harness it in support of a general model for combining beliefs, as we do in Section (4), we first need to develop a basic language for describing the magnitude of evidence reflected by one's credence function. This is a dimension of the person's doxastic state that is not captured by the credence profile. Accordingly, we will describe below a natural Bayesian approach for modeling exchangeable data which will allow us to explain these ideas more clearly. While the next two subsections may seem unduly technical, we spell things out carefully because doing so will allow us to substantially simplify the core material in Sections (4)-(6).

3.1 Characterizing Resilience

Suppose again we have two people, Ahmed and Beatrice. They will each draw n balls from an urn with replacement. The urn contains white and blue balls, with θ as the unknown proportion of white balls. In a draw of n balls, let r be the number of white balls and, hence, $n - r$ the number of blue balls. Each person's credences may be about θ, or they may be about the probability that the next ball to be drawn will be white or blue – i.e., predictions on $\widetilde{X}$, where $\widetilde{X} = 1$ represents a white ball and $\widetilde{X} = 0$ represents a blue ball.

Predicting the next ball and estimating θ both correspond to different practical problems. For example, in the context of Covid-19, a doctor might be interested in the probability that the next patient she sees is positive (predicting the color of the next ball) whereas a policymaker in her city may be more interested in the proportion of the population that is positive (estimating θ). As long as one uses a proper scoring rule, the Bayesian logic follows a similar structure for either task.

Accordingly, it is not enough to have a "credence" over the space of outcomes because there are many different distributions for θ which correspond to the same predictions about which ball will be drawn. Thus, we first need to specify a full distribution for θ. Once we have that, we can make point estimates, interval estimates, and predictions about $\widetilde{X}$.

In a draw of n balls from an urn that contains only white and blue balls, the data are generated according to a Bernoulli process with the following likelihood function:

$$f(r|\theta, n) = \theta^r(1-\theta)^{n-r}. \tag{1}$$

Now we need to identify prior beliefs regarding θ. In the Bayesian approach, a good candidate prior for θ when data is generated according to (1) is the so-called beta distribution, because it is a very flexible distribution with two parameters, $\alpha \geq 0$ and $\beta \geq 0$, accommodating a wide variety of information states regarding a Bernoulli process and it arises naturally in the context of modeling binary exchangeable data (Lindley and Phillips, 1976). An early development of this model can be found in Johnson (1924), It was later applied by Carnap (Carnap, 1950, 1952), in his construction of the continuum of inductive methods, and by DeFinetti (De Finetti, 1937), in his refinement of Laplace's Rule of Succession.

Let $\pi(\theta)$ be the prior probability density for θ, where

$$\pi(\theta) = f(\theta|\alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{2}$$

is a beta density function with $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ and $\Gamma(n) = (n-1)!$. The core of this distribution, its kernel, is given by $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. The combinatorial term in front is a normalizing constant. The mean of a beta distribution is given by $\mathrm{E}[\theta] = \alpha/(\alpha+\beta)$. This will be an important quantity in the material to follow, as will α and β. With the likelihood in (1) and the prior in (2), the posterior density for θ is

$$\pi(\theta|r, n) = f(\theta|\alpha + r, \beta + n - r) \propto \theta^{\alpha+r-1}(1-\theta)^{\beta+n-r-1}. \tag{3}$$

The posterior distribution is of the same kernel form as the prior distribution. This is because a beta distribution is *conjugate* to the Bernoulli

process. This means that if we start with a beta prior for θ, and update via Bayes' Rule with data from a Bernoulli process, our posterior will likewise be beta but with updated parameters.

Such a model lends itself to a very intuitive interpretation.[6] The parameters of the updated beta distribution (the posterior distribution) are given by the sum of α and the number of white balls, r, together with the sum of β and the number of blue balls, $n - r$. As a result, the parameters α and β can be interpreted as pseudo observations or pseudo counts upon which the prior beliefs are based. For instance, to say that one has a beta(2, 2) prior for the proportion θ of white balls in the urn is equivalent to assuming that prior to making the actual draws, that person observed two balls of each color.

Bayesian updating is very simple and intuitive within a beta-Bernoulli model. If we start with a beta(7, 3) prior, and we observe 4 out of 10 white balls, our posterior for θ would be beta(7+4, 3+6). A beta(1, 1) prior for θ is the uniform or flat prior. This would also be the maximum entropy prior for a proportion.

If we want to formulate a credence about $\widetilde{X}$ (the color of the next ball to be drawn), we need the predictive distribution. Assuming that the draws are conditionally independent given θ, this is given by:

$$\begin{aligned} \mathrm{P}(\widetilde{X} = 1|r,n) &= \int_0^1 \mathrm{P}(\widetilde{X} = 1|\theta)\pi(\theta|r,n)d\theta \\ &= \mathrm{E}[\theta|r,n] = \frac{\alpha + r}{\alpha + \beta + n}. \end{aligned} \tag{4}$$

Huttegger (2017a,b) refers to this expression as the Generalized Rule of Succession and shows that this form of the predictive probability follows from several modest assumptions about the structure of the data-generating process, in particular exchangeability, which will be satisfied throughout. From a decision-theoretic perspective, the posterior mean minimizes expected square error loss. The important point for us is that when the problem is fully specified, each person will have a full distribution about θ. They will then base their prediction on the mean of that distribution.

We can now put this model to its first use and capture the notion of resilience for a credence function. Suppose A starts with a beta(1,1) prior and B starts with a beta(10, 10) prior regarding θ, the proportion of white balls in the urn. They both obtain equivalent non overlapping evidence: namely, each draws 10 balls, 6 of which are white. Using (3), we can determine that their posteriors will become beta(7, 5) and beta(16, 14), respectively. Using (4), we conclude that A's probability that the next ball to be drawn is white moves from 0.5 to 0.58 whereas B's moves from 0.5 to 0.53. B is much more resolute in her prior, even though the actual credal value (the valence) is the same among them. The resilience of a credence function for θ, therefore, corresponds to the size of α and β.

> **Resilience**. Let $(\alpha + \beta)_A$ denote the size of the sum of α and β in A's beta distribution for θ. If $(\alpha + \beta)_A > (\alpha + \beta)_B$ then A's distribution for θ is more resilient.

The higher $(\alpha + \beta)$ the more resolute the person will be about her credences. Keep in mind that after we make observations, r contributes to our new α, $n - r$ contributes to our new β, and n contributes to $\alpha + \beta$. As a result, the preceding definition captures Joyce's dictum that resilience corresponds to the weight of one's data – i.e., to n.

As the above example comparing a beta(1, 1) prior against a beta(10, 10) prior illustrates, it is possible to have equal valence and unequal resilience. This is why the credence profile sufficiency assumption is problematic. When we combine valences, there are many degrees of resilience compatible with the ensuing group credence. Which level of resilience we then impute to the group will later affect how it responds to new evidence. As a result, credence is not reducible to valence.

6. See also Babic et al. (2024) for a discussion of the beta Bernoulli model and its interpretation.

Without capturing resilience we fail to specify every individual's full quantification of uncertainty.

3.2 *Synchronic Resilience*

One might wonder whether the only way resilience reveals itself is under learning experiences – diachronically, so to speak. This is not the case. We can illustrate the difference between resolute and irresolute agents synchronically as well.

The point estimate for θ, $\hat{\theta}$, is the same as the predictive probability for a single draw, $P(\widetilde{X} = 1)$. But instead of just producing the value we think is most likely – which we know to be false anyway, since θ is continuous – we can instead produce an interval estimate for θ. In Bayesian inference, a $(1-\gamma)100\%$ credible interval (a, b) satisfies:

$$P(a < \theta < b|r, n) = \int_a^b \pi(\theta|r, n)d\theta = 1 - \gamma. \tag{5}$$

Consider an example: A's credences are beta(2, 2) whereas B's credences are beta(100, 100). Both estimate the proportion of white balls to be 0.5, and both estimate that the probability of the next ball being white is 0.5. Their probabilities (valences) are identical, as are their predictions about the next ball. However, A's 95% credible interval is $(0.1, 0.9)$ whereas B's 95% credible interval is $(0.43, 0.57)$. This is a dramatic difference in uncertainty around the prediction. A is very open minded, whereas B is quite dogmatic.

Indeed, diachronic and synchronic resilience are related. As α and β increase, the variance of the distribution, given here by $\sigma^2 = \alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$, decreases. This is to be expected – intuitively, large n implies that it is hard to change the distribution with extra information (diachronic), while small σ^2 implies that the current estimate is tight (synchronic). So when we increase α and β we tighten the variance. Therefore, for a given mean (point estimate) by increasing α and β we ordinarily shrink the width of the credible interval. This is easiest to illustrate if we approximate a beta prior with a normal distribution,[7] where intervals are symmetric. In this case, a 95% credible interval simplifies to $\mu \pm 1.96\sigma$, and in our example,

$$\mu = \alpha/(\alpha + \beta), \text{ and}$$
$$\sigma = (\alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1)))^{1/2}.$$

We can see that as α and β increase, then σ decreases and so for a given μ, the length of the credible interval shrinks.

4. The Evidence-First Method

We now present the evidence-first method. In Section (4.1), we describe the approach informally when there is no shared evidence. In Section (4.2), we develop the idea mathematically for the general case where evidence is overlapping. In Sections (4.3)-(4.4), we show that in the special case with minimal overlapping evidence, we recover the ordinary (weighted) averaging rule. And in the special case with a uniform prior and no overlap, we recover the rule articulated in Easwaran et al. (2016), which they call Upco (Section 4.5).

4.1 *A Simple Example*

Suppose we have an urn with blue and white balls in unknown proportion θ, and two people, A and B, each of whom holds a uniform beta(1, 1) prior for θ. They each draw 10 balls, independently, and observe 4 and 7 white balls, respectively, with no overlap. What should the group distribution be?

First, since both A and B approach the problem with a uniform prior, the group prior before any observations are made should be uniform as well.[8] Second, and more importantly, we have to make explicit the

7. This approximation follows from the Central Limit Theorem and is reasonably accurate if $\alpha > 5$ and $\beta > 5$.
8. We explain and argue for this in more detail in Section (4.2), below. For now the exposition is informal to motivate the reader's intuition.

group's shared evidence. We have 4+7=11 distinct white balls and 6+3=9 distinct blue balls, for a total of 20 balls. We can think of the group as accomplishing a division of labor – assigning 10 draws for A to handle, and 10 draws for B to handle. They each do their job and come back to combine the evidence. The group distribution is therefore beta(12, 10). This is a full distribution for the unknown proportion θ, from which we can derive any statistic of interest. For instance, the group's (posterior mean) point estimate for θ is $12/22 = 0.54$. The median is 0.55. The 95% credible interval is $(0.34, 0.74)$. We now have a full representation of the group's uncertainty.

By contrast, if we combine credences through simple ordinary averaging, for instance, we might pool the two point estimates from A's beta(5, 7) and B's beta(8, 4) distribution, which would be $(0.41 + 0.66)/2 = 0.53$. But notice that on ordinary averaging approaches we do not have the full distribution, using instead only the probabilities as described by the credence profile. As a result, it would be impossible to determine whether person A's 0.41 estimate came from a beta(5, 7) distribution, a beta(10, 14) distribution, a beta(20, 28) distribution, or any other beta distribution which satisfies $\alpha/(\alpha + \beta) = 0.41$. Because all of these distributions are compatible with the reported probability, we also cannot say how the group should update if it makes additional observations. For example, if its prior distribution is beta(5, 7) and it observes 2 white balls it should move to beta(7,7) and a 0.5 estimate of θ. But if its prior distribution is beta(20, 28) and it observes 2 white balls then it should move to beta(22,28) and an estimate of 0.44 for θ. Likewise, the 95% credible interval in the beta(5, 7) case is (0.17, 0.69) whereas in the beta(20, 28) case it is (0.28, 0.55).

Further, it is well known that ordinary averaging is not commutative with respect to updating: updating and combining does not always give the same result as combining then updating.[9] Our approach, by comparison, does not have this problem, and we establish and discuss this for the general case in Theorem 1 below. Indeed, it is easy to see that this will be true because we are simply summing up the number of observations in each category. Since addition is commutative, so too is the evidence-first method.

4.2 *The Core Idea*

The preceding examples are particularly easy to handle because each person receives independent signals – no balls are observed together. But it is rarely the case that a group of people approach a problem with mutually exclusive private information. When some balls are observed in common, the key is to capture overlapping evidence appropriately.[10] This way, everyone contributes exactly their evidence, and only their evidence, to the group belief. We now generalize the above idea and mathematically formulate the evidence-first method. This model captures both dependence and resilience.

We use the case of two people and two categories for maximal simplicity. Extending to n people and k categories is straightforward; but since the number of parameters grows quickly, in both k and n, it risks burying our message in the details. In a problem with two categories (two colors of balls), for two people, we need to know six quantities/parameters: the number of white balls each observed, the number of blue balls each observed, and the number of each color of balls observed in common.

Suppose we have two people, $i = 1, 2$, who will estimate the probability p that the next ball drawn from the urn will be white (label it as a success). Each person has her own full subjective distribution over p, which is a beta distribution with parameters r_i and $n_i - r_i$:

$$\pi_i(p) \propto p^{r_i - 1}(1 - p)^{n_i - r_i - 1}. \tag{6}$$

9. Russell et al. (2015) (Fact 4), Dietrich (2019) (Theorem 2), and Pettigrew (2019) (Theorem 3).

10. Clemen (1987) proposes a similar approach but we expand on this work in several directions: by proving that the method is commutative under updating, by explaining when it is equivalent to weighted averaging, and by connecting it to likelihoodist approaches in philosophy.

Thus, for person i, the probability that the next ball will be white corresponds to $\mathrm{E}_i[p] = \frac{r_i}{n_i}$. Moreover, n_i captures the notion of resilience described above – the larger n_i, the more resolute person i is that the probability is close to p_i.

In order to combine these two people's opinions/credences, we now need to model their shared information structure – which must reflect the way each came to their subjective probability distributions and any overlap in their evidence. Our model is as follows: Every person starts with a beta prior with parameters α_0 and β_0. Typically, these parameters will be small, like $0 \leq \alpha_0 \leq 1$ and $0 \leq \beta_0 \leq 1$. Such a prior is proper (i.e., there exists a normalizing constant) if $\alpha_0 > 0$ and $\beta_0 > 0$. This is often, but not always, the case. We will consider some improper priors below.

Each person will observe a few draws from this urn. This is their evidence. More specifically, α_c is the number of successes observed by both people, β_c is the number of failures observed by both people, α_i is the number of successes observed only by person i, and β_i is the number of failures observed only by person i. Now we can specify r_i and n_i. In particular, $r_i = \alpha_i + \alpha_c + \alpha_0$ and $n_i - r_i = \beta_i + \beta_c + \beta_0$.

The group credence function, which we will denote by $\Pi(p)$ (i.e., small π for the individual distribution, large Π for the group distribution), then corresponds to a situation when all of this information is combined. Therefore, it is a beta distribution with parameters $r^* = r_1 + r_2 - \alpha_c - \alpha_0$ and $n^* - r^* = (n_1 - r_1) + (n_2 - r_2) - \beta_c - \beta_0$:

$$\Pi(p|r,n) \propto p^{r^*-1}(1-p)^{n-r^*-1}. \tag{7}$$

This implies that the group probability is $p^* = \frac{r^*}{n^*}$ and the new sample size (resilience) of the group is equal to n^*. The group probability p^* can be expressed (using $r_i = p_i n_i$ and $n^* = n_1 + n_2 - \alpha_c - \alpha_0 - \beta_c - \beta_0$) as

$$p^* = \frac{r^*}{n^*} = \frac{p_1 n_1 + p_2 n_2 - \alpha_c - \alpha_0}{n_1 + n_2 - \alpha_c - \alpha_0 - \beta_c - \beta_0}. \tag{8}$$

This final quantity, in (8), is what the group uses as its probabilistic estimate. This is the reported probability – the valence. However, unlike approaches which focus on deriving the probability alone, we derive the full group distribution, (7), and the two are related because $p^* = \mathrm{E}[p]$. In sum, equations (6)-(8) describe the evidence-first method. We now state a useful theorem about this method.

> **Theorem 1 (Update Commutativity)**. Let $\pi_i(p)$ be i's prior distribution for p, for $i = 1, 2$. Let $\Pi(p)$ be the group prior, derived using (7). Let $\pi_i(p|r,n)$ be i's posterior distribution for p, obtained from $\pi_i(p)$ via Bayes' Rule, after learning new information r and $n - r$, and let $\Pi(p|r,n)$ be the group posterior, obtained from $\Pi(p)$ via Bayes' Rule, also after learning r and $n - r$. Finally, with slight abuse of notation, let $\Pi(p)|r,n$ be the group posterior obtained if we first update $\pi_i(p)$ to $\pi_i(p|r,n)$ and then combine $\pi_i(p|r,n)$ using (7). Then,
>
> $$\Pi(p)|r,n = \Pi(p|r,n). \tag{9}$$
>
> Proof in the Appendix.

Thus, unlike ordinary averaging (weighted or simple) our approach is update commutative. The proof of this theorem is straightforward. Any distribution in our framework is simply characterized by the number of successes and failures, with probability given by the proportion of successes. So when we combine the distributions, we just count the total number of successes and failures. The only wrinkle to keep in mind is that we must avoid double counting trials that were observed by both people. When we "combine and then update" we first count the total number of successes and failures in the priors, and then add new successes and failures. When we "update and then combine", we first count successes and failures for each agent, and then count the total.

There is another aspect of the model worth flagging. We suppose our individuals have a common diffuse/uniform prior, since we use

α_0 and β_0 instead of α_0^i and β_0^i. This assumption is not essential and it can ultimately be dropped, but because some readers may find it problematic, we explain this modeling choice. To understand why we make this assumption, consider two different cases. First, suppose A and B have no prior information, and they adopt a uniform distribution on the basis of something like the Laplacean principle of indifference.[11] That is, both are completely ignorant before making the relevant observations and their prior is a true flat ur-prior.[12] Thus, $\alpha_0 = 1$ and $\beta_0 = 1$. Suppose we now want to combine these priors before updating on any information. What should the group distribution be? Our model implies that it should be beta$(1, 1)$ and not beta$(2, 2)$. This is intentional. We do not want two truly ignorant individual priors to sum up to an informative group prior. Another way to put this is that combining ignorance with ignorance should not lead to wisdom or confidence, just as 0 + 0 = 0.

Second, suppose A and B do have concrete prior information. For example, A has previously drawn balls from an urn in a game of chance at the Ringling Brothers circus whereas B has done so at the Barnum & Bailey Circus. They now find themselves together at the Ramos Brothers Circus, having to make predictions about an urn neither has previously encountered. But they happen to know that all three circuses keep a similar house edge so the proportions cannot vary too widely. Suppose they start with beta(7, 3) and beta(3, 7) priors for the proportion of white balls in the urn at the Ramos Brothers Circus. They then observe 4 draws, two of which are white and two of which are blue. What should the group distribution be?

To answer this, we must make clear the sequence of updating. What happened here is that both people updated on two sets of observations/ two experiments – first, independently, at the Ringling Brothers / Barnum & Bailey Circus, and second, at the Ramos Brothers Circus, together. Thus, we need to determine what their ur-prior was before both sets of observations. Suppose again it was beta(1, 1).[13] This implies that they each observed 8 balls at the first circus, which is why the sum of α and β for both is 10 before the second circus.

Which approach they use to set their ur-prior does not matter as long as there is a shared understanding of what rationality calls for in the absence of information. Accordingly, our assumption is like a weak version of the common prior familiar from microeconomic theory.[14] It is "weak" in the sense that we do not assume rationality writ-large requires universal agreement about ur-priors. We simply assume that the members of the group agree in this regard. Importantly, however, they can pick any starting point and our assumption that a uniform prior corresponds to an ignorant or uninformative distribution is merely illustrative.

Given this specification of the problem – beta(1, 1) ur-priors, followed by (6,2) and (2,6) white/blue observations alone at the first circus, followed by (2, 2) white/blue observations at the Ramos Brothers Circus – the group posterior distribution becomes beta(11, 11). We subtract only the initial $\alpha_0 = \beta_0 = 1$ from the ur-prior and *not* the (6, 2) / (2,6) observations made at the first circus, since these are ordinary independent observations. If they started with beta(0, 0) ur-priors, the

11. Laplace (1814). For recent defenses, see White (2010) and Pettigrew (2019). For the case of a proportion, the flat prior is also the maximum entropy prior (Jaynes, 1957a,b).
12. By ur-prior, we refer to the stylized prior that an agent may hold before observing any evidence whatsoever – the Lewisian superbaby (Hájek, ms).
13. The flat beta(1, 1) prior is merely illustrative, though Babic (2019) argues it can be considered maximally safe under certain loss functions. It may be instead that they adopted maximally ignorant beta(0, 0) distributions, the so-called Haldane priors (Robert, 2007). Or perhaps due to symmetry considerations, such as those articulated in Zabell (2005), they adopt the invariant Jeffreys' prior, which in this case corresponds to a beta(1/2, 1/2) distribution (Jeffreys, 1946). Notice that all the above methods agree on one thing: namely, that α_0 and β_0 are very small, and in all three cases just a little bit of information leads to similar predictions. Our approach is compatible with any assumption one wants to make about how to represent true ignorance, as long as one is clear about that assumption so that we know which part of their distribution is informed by the evidence, and which part is informed by their prior commitments.
14. This assumption is most notably associated with Harsanyi (1987) and Aumann (1987).

group posterior would be beta(12, 12). The message is that we must be clear both about how the prior is selected before observations are made, and about what evidence is available to each person, both individually and jointly. In short, the only burden our framework imposes is that when modeling common information, we have to be careful to model it via α_c and β_c and not α_0 and β_0.[15] To further illuminate our model, we will look at several cases where (8) takes a simple form.

4.3 *Limited Evidential Overlap*

In our approach, the simple case where there is no evidential overlap corresponds to what Dietrich and Spiekermann (2013) would describe as a set of opinions which are common cause conditionally independent. Let us examine this kind of situation. Suppose

$$\frac{\alpha_c + \alpha_0 + \beta_c + \beta_0}{n_1 + n_2} \ll 1, \qquad \frac{\alpha_c + \alpha_0}{p_1 n_1 + p_2 n_2} \ll 1. \tag{10}$$

This is the case if both people start with completely ignorant beta(0, 0) priors and observe no information in common. That is, $\alpha_c = \alpha_0 = \beta_c = \beta_0 = 0$. Then, from (8),

$$\begin{aligned} p^* &= \frac{p_1 \frac{n_1}{n_1+n_2} + p_2 \frac{n_2}{n_1+n_2} - \frac{\alpha_c+\alpha_0}{n_1+n_2}}{1 - \frac{\alpha_c+\alpha_0+\beta_c+\beta_0}{n_1+n_2}} \\ &= \left(p_1 \frac{n_1}{n_1+n_2} + p_2 \frac{n_2}{n_1+n_2}\right)\left(1 - \frac{\alpha_c+\alpha_0}{p_1 n_1 + p_2 n_2}\right) \\ &\quad \left(1 - \frac{\alpha_c+\alpha_0+\beta_c+\beta_0}{n_1+n_2}\right)^{-1} \\ &\approx p_1 \frac{n_1}{n_1+n_2} + p_2 \frac{n_2}{n_1+n_2}. \end{aligned} \tag{11}$$

In this case, we recover the ordinary weighted averaging rule, as defended in Moss (2011) and (Pettigrew, 2019), among others, where the weights are determined by n_1 and n_2, the total number of each person's observations – i.e., their resilience. This is intuitive, and indeed consistent with Pettigrew (2019)'s defense of ordinary weighted averaging because the more resolute of the two people will exert a greater weight on the group credence function. As Pettigrew suggests, it appears reasonable that the weights of an aggregation function reflect expertise – so that more knowledgeable members exert more influence on the group's belief. It is also consistent with the interpretation given to the weights in ordinary weighted averaging in Romeijn (2024). Romeijn interprets the weights in terms of the truth conduciveness that one agent assigns to the other, which can also be thought of in terms of the trust placed in them.[16] Accordingly, not only do we recover the ordinary weighted averaging rule, but in doing so we also provide a principled reason for

15. In this sense, our weak common prior assumption might be described as a local or group-level impermissivism about the requirements of rationality with respect to an ur-prior. While one can find many defenses of both objectivism in the selection of priors (e.g., Williamson (2010)) and uniqueness at large (such as Greco and Hedden (2016)), we do not need to assume such a strong position, as even the weak/local impermissivist assumption is ultimately a modeling choice and may be relaxed.

16. Truth conduciveness, following its meaning in the Condorcet jury theorems, implies that it is more probable that the person believes (i.e. 'votes for') a proposition if it is true than if it is false. See Romeijn and Atkinson (2011).

how to assign the weights in that rule: namely, by using them to encode resilience.

4.4 *Equal Resilience*

Consider the case where $n_1 = n_2 = n$. Here things become even more straightforward since if common information is small, then we will combine individual credences by simple ordinary averaging:

$$p^* = \frac{p_1 + p_2}{2}. \tag{12}$$

This is intuitive. When resilience is equal, the weights in the ordinary averaging rule ought to be equal. And it can be motivated on similar grounds as above: if we have a group of equally knowledgeable agents, it is reasonable to assign them equal weights.

But according to (8), the prior parameters (α_0 and β_0) and the number of successes and failures observed by both people (α_c and β_c) can change this formula. From (8),

$$p^* = \frac{(p_1 + p_2)\, n - \alpha_c - \alpha_0}{2n - \alpha_c - \alpha_0 - \beta_c - \beta_0}. \tag{13}$$

Without loss of generality, suppose that $p_1 \leq p_2$. Then $0 \leq \alpha_c + \alpha_0 \leq p_1 n$ and $0 \leq \beta_c + \beta_0 \leq (1 - p_2)\, n$. The case where $\alpha_c + \alpha_0 = 0$, $\beta_c + \beta_0 = 0$ corresponds to a situation where the body of common evidence is small, and as a result, $p^* = \frac{p_1 + p_2}{2}$, $n^* = 2n$. But now consider two further cases, where either $\alpha_c + \alpha_0$ or $\beta_c + \beta_0$ is at a maximum or a minimum.

Case 1. $\alpha_c + \alpha_0 = p_1 n$, $\beta_c + \beta_0 = 0$. Then, by (13), $p^* = \frac{p_2 n}{2n - p_1 n} = \frac{p_2}{2 - p_1}$; $n^* = (2 - p_1)\, n$. So not only is the combined resilience now less than $2n$, but p^* is different from the average of individual probabilities. Even if $p_1 = p_2 = p$, the combined probability is still $p^* = \frac{p}{2-p} < p$. This is because in this case successes are observed by both people together, while failures are observed by each person separately.

Case 2. $\alpha_c + \alpha_0 = 0$, $\beta_c + \beta_0 = (1 - p_2)\, n$. Then, by (13), $p^* = \frac{(p_1 + p_2) n}{2n - (1 - p_2) n} = \frac{p_1 + p_2}{1 + p_2}$; $n^* = (1 + p_2)\, n$. As in the previous case, the combined probability is again different from what simple averaging would suggest.

4.5 *A Closer Look at Priors*

We now examine the impact of the prior distributions. Assume $\alpha_c = \beta_c = 0$ and $\alpha_0 = \beta_0 = d$. Let $p_a = \frac{p_1 + p_2}{2}$, which would be the combined probability under ordinary simple averaging. Then,

$$\begin{aligned} p^* &= \frac{(p_1 + p_2)\, n - d}{2n - 2d} \\ &= \frac{\frac{p_1 + p_2}{2} - \frac{d}{2n}}{1 - \frac{d}{n}} \\ &= \frac{p_a \left(1 - \frac{d}{n}\right) + p_a \frac{d}{n} - \frac{d}{2n}}{1 - \frac{d}{n}} \\ &= p_a + \left(p_a - \frac{1}{2}\right) \frac{d}{n} \frac{n}{n - d} \\ &= p_a + \left(p_a - \frac{1}{2}\right) \frac{d}{n - d}. \end{aligned} \tag{14}$$

This highlights an important feature of our model, which we call extremization (Lichtendahl et al., 2021). By extremization we refer to a phenomenon that Easwaran et al. (2016) call synergy. It suggests that the group belief can lie outside the interval formed by the lower and upper bounds of individual beliefs. Examining the last line in (14), we can see that p^* extremizes away from $\frac{1}{2}$ whenever the quantity on the right side of the sum is not 0. That is, the group credence extremizes

unless $d = 0$, or $p_a = \frac{1}{2}$, or $r_i = 0$ or $n_i - r_i = 0$. Meanwhile, adopting a uniform prior in the above case would correspond to a situation where $d = 1$, and for small n the extremization can be quite substantial. Note also that extremization will occur even if $p_1 = p_2 = p_a$.

Extremization is not possible under ordinary averaging rules, where the group belief must lie in the convex hull of the set of individual beliefs. But we agree with Easwaran et al. (2016) that extremizing can be rational, especially in cases where, as here, the common evidence is small. Consider a more realistic scenario. A company's executive committee is predicting whether the company will break even next year. It consists of the heads of marketing, finance, and operations. All three independently report that the company has a 97% probability of breaking even. Given that each of these executives is coming from a different area of the company, and is likely basing their forecast on largely independent evidence, it is particularly plausible that the group credence should be above 0.97. Indeed, if the credence remains at 0.97, as ordinary averaging requires, we are likely throwing away information (see also Christensen, 2011).

Finally, note that under a uniform prior, i.e., where $\alpha_0 = \beta_0 = d = 1$, the posterior distribution is proportional to the likelihood. Therefore, if we combine two distributions, and we assume that each person started with a uniform prior and received independent signals, i.e., $\alpha_c = \beta_c = 0$, then the combined posterior will be proportional to the product of their individual distributions. In such a case, the individual distribution of person i is beta with r_i and $n_i - r_i$, and the combined distribution is beta with $r_1 + r_2 - 1$ and $n_1 - r_1 + n_2 - r_2 - 1$, which is what we would get if we multiply their individual distributions:

$$\begin{aligned} &p^{r_1-1}(1-p)^{n_1-r_1-1}p^{r_2-1}(1-p)^{n_2-r_2-1} \\ &= p^{r_1+r_2-1-1}(1-p)^{n_1-r_1+n_2-r_2-1-1}. \end{aligned} \tag{15}$$

Therefore, with independent signals under uniform priors we recover the so-called Upco rule from Easwaran et al. (2016) for updating on the credences of others. Upco is derived as the product of odds ratios – for one person, the odds ratio is $p/(1-p)$, for another it is $q/(1-q)$, so the product is $qp/[(1-p)(1-q)]$, and after normalization we obtain Upco.

But, our method produces a combined distribution for p, and the expected value of that distribution would be the probability for the next ball drawn to be white. In Upco on the other hand (as defined on Easwaran et al., 2016, pg. 3), the rule applies directly to the probabilities of the next ball, which are not sensitive to considerations of resilience. Thus our approach coincides with Upco only under uniform priors and independent signals. For example, if $r_1 = r_2 = 100$ and $n_1 = n_2 = 1000$, then our combined probability will still be around 10%; with Upco, if we combine $p = q = 10\%$ we would get a group probability of about 1%.

5. A Worked Example: Hiring a Netflix Developer

To get a better feel for the evidence-first method, consider an extended and more realistic example.

> **Netflix**. Netflix is interested in hiring an original series developer. This will be a full-time employee whose job is to bring new pitches, specs, etc., to the streaming service. The search committee consists of two members, Ahmed and Beatrice. The shortlist of competing candidates are all individuals with significant prior experience in developing shows. The committee considers a show successful if it runs for one season or more and turns a profit. Of interest, then, is the developer's probability of creating a successful show. They are now considering a well-known developer named Jean Marscome.

Suppose A and B each report the following probabilities of JM's success: 0.7 and 0.3, respectively. These are their naked probabilities – or valences – and we now know that this is not enough to appropriately combine their beliefs. Rather, we should elicit the members' individual evidence

on which these predictions are based and piece together their full distributions, from which we then derive the group distribution and make predictions about JM.

Thus we first need to know their priors for p. Suppose A and B agree that in the absence of information one should apply the principle of indifference and they both adopted a uniform prior. In fact, suppose that this is standard Netflix company policy in the context of recruiting: before any information on a candidate is obtained, the hiring committee must treat the candidate's probability of success as uniform on $[0, 1]$. This is not an unwise policy – it may be enforced to avoid favoritism among job candidates.

Next, suppose they each lay their cards on the table. A is familiar with 8 of JM's shows, 6 of which were successful, and 2 of which were unsuccessful. Meanwhile, B is also familiar with 8 of JM's shows, but 2 of them were successful and 6 of them were unsuccessful. We can now account for the resilience of their credences, because we have the weight of their evidence. But we still need to untangle potential dependencies.

Finally, A and B list the JM shows they are familiar with, identifying each as a success or failure. As part of this exercise, we learn that they have 2 shows in common, both of which were a failure. We now know all six parameters $(\alpha_0, \alpha_i, \alpha_c, \beta_0, \beta_i, \beta_c)$. From these, we can determine n and r, compute the individual distributions, and predictions, and the group distribution and prediction. Table (1) summarizes the above.

	Ahmed	Beatrice	Group (Netflix)
α_0	1	1	1
β_0	1	1	1
α_i	6	2	8
β_i	0	4	4
α_c	0	0	0
β_c	2	2	2
n	8	8	$8+8-2=14$
$r = \alpha_0 + \alpha_i + \alpha_c$	$1+6+0=7$	$1+2+0=3$	$1+8+0=9$
$n - r = \beta_0 + \beta_i + \beta_c$	$1+0+2=3$	$1+4+2=7$	$1+4+2=7$
Full distribution for p	beta$(7,3)$	beta$(3,7)$	beta$(9,7)$
Prediction of JM's success, i.e., E$[p]$	$7/10=0.7$	$3/10=0.3$	$9/16=0.5625$

Table 1: Individual and group beliefs in Netflix example.

We highlight several points. First, the group prediction is not a simple average of the individual predictions. It is tilted upward because there are overlapping failures and no overlapping successes. By comparison, Russell et al. (2015) and Dietrich (2019)'s geometric mean would produce a prediction of 0.46 without normalization, since $(0.7 \times 0.3)^{1/2} = 0.46$, putting more weight on Beatrice, and 0.5 with normalization.

Notice, also, the effect of resilience. If after combining their beliefs into a group distribution, they were to watch three of JM's shows together, all of them a failure, they would update to a group distribution of beta(9, 10) and the prediction of JM's success would then drop from 0.56 to 0.47. Now suppose we double all the values in the original example, so that the group belief is beta(18, 14) before they watch any shows together. This time, again, they watch three shows, all failures, thereby updating the group distribution to beta(18, 17). Now the prediction drops from 0.56 to 0.51. Because such a group is more resolute in its estimate of JM's success, it responds less to 3 failures than it did in the original case. This is a facet of the situation that the current approaches in the literature are not equipped to handle.

Finally, our approach streamlines certain distinctions often made in the aggregation literature. Dietrich (2019), for example, argues that there are "different types of group Bayesianism, depending on the kind of information on which one requires groups to conditionalize [public, semi private, private]" (pg. 721). This tri-partite distinction is a necessary byproduct of the assumption that the credence profile is a sufficient summary statistic of individual beliefs. Our approach, by comparison, does not require a multitude of Bayesianisms. There is only one way to be Bayesian, namely, by passing whatever is learnt through Bayes' Rule. Public information consists of balls observed by every member of the group. Semi private information consists of balls observed by two or more but not all members of the group. Private information consists of balls observed by only one member of the group. To further drive the reader's intuition, we include in Figure (1) a visual representation of the Netflix situation.

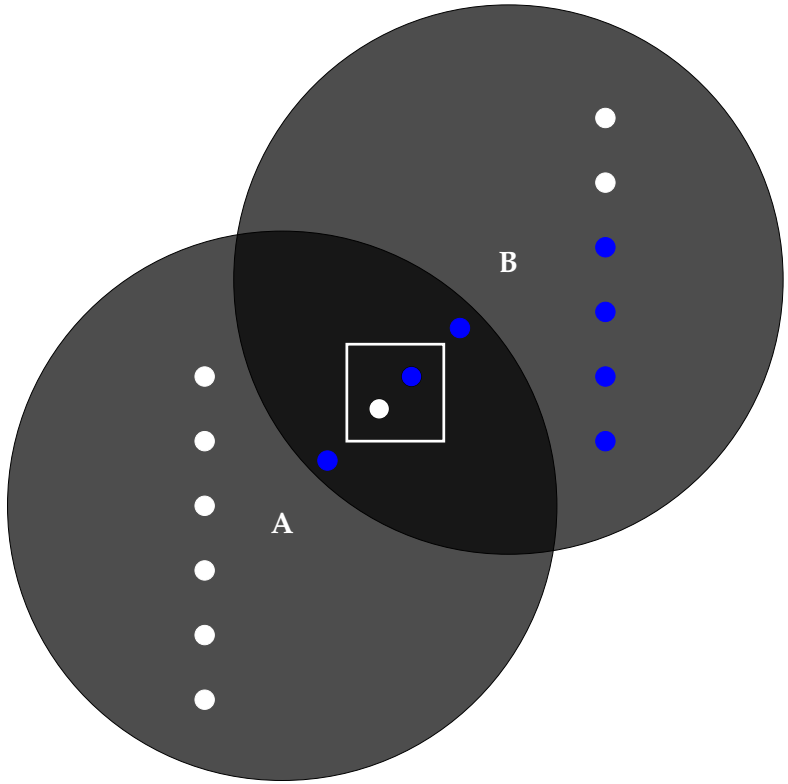

Figure 1: A and B's combined credences about JM. The white balls are successes and the blue balls are failures. The points inside the box are the common uniform priors. The points outside the box but inside the intersection constitute shared evidence. And the points outside the intersection constitute each person's individual evidence.

6. Scope and Applicability

One might wonder whether the range of aggregation problems within the scope of our prescriptive approach is too limited. As in the Netflix example, our framework may at first blush seem to call for conditions

that are often unmet in real life: namely, the individuals in the group should be able to have a conversation and reveal their total evidence. But what about cases where individual credences do not arise from such a clean and well-specified process?

While we think the requisite conditions are not as unattainable as may first appear, it is nonetheless true that sometimes we cannot so neatly disclose our evidence. Indeed there may be times where a decision maker is faced with the unenviable task of combining bare individual forecasts (which may have been compiled for them by someone else, or made a long time ago, or by experts who are now inaccessible, etc.). In short, one might suggest that the credence profile sufficiency assumption, common in the literature, is not so much a desideratum of the belief combination problem as it is a description of the hard reality in which beliefs must be combined.[17]

Even in such cases, however, our approach remains valuable as a normative benchmark for evaluating the rationality of group beliefs. To understand how, suppose we really are in a situation where we have to aggregate credences without access to the full evidence set, or perhaps to any evidence at all. In such cases, we can apply the framework we suggest in reverse. Instead of using this approach as a recipe for how to reach a specific group distribution, we can instead identify a range of permissible group credences which are consistent with our best estimate about the plausible evidence histories of the individual members.

For example, suppose that in the case of drawing marbles from urns, we have A and B's predictions that the next marble will be white. We also know that they observed some of the same draws, but we are not sure how many they observed in common. Suppose that we have no further information. In this situation, we have to make some assumptions about their resiliences, and about the extent of their evidential overlap. There are many evidence histories compatible with their predictions. Accordingly, we can construct upper and lower bounds on what the rationally permissible group prediction should be.

17. Thanks to Frederick Eberhardt for raising this consideration.

The normative guidance that our approach produces in this case is not as specific as in the Netflix example, but that is to be expected since the information structure is now far more impoverished.

To make this more precise, consider how we would make such evaluative judgments without knowing the actual evidence histories. First, we need to estimate individual resiliences, giving more weight to sharper distributions. If we have no basis on which to estimate these, we might start by assuming that everyone's resiliences are the same (for similar reasons that would motivate the Laplacean Principle of Indifference). Next, we have to estimate the overlap in their evidence. This will depend on our assessment of the number of evidence histories consistent with the individual credences. But helpfully, our results from Section (4) provide some general bounds on what is permissible.

If we go back to our Cases 1 and 2 (Section 4.4), we have there equal resilience leading to simple averaging, such that $p^* = (p_1 + p_2)/2$, but only if common information is small. Thus we can now reconsider extreme cases of overlap to see what happens. Assuming again without loss of generality that $p_1 < p_2$, then in Case 1 α_c is at a maximum and by (13), $p^* = p_2/(2 - p_1)$. In Case 2, β_c is at a maximum and by (13), $p^* = (p_1 + p_2)/(1 + p_2)$. This means that

$$\frac{p_2}{2 - p_1} < p^* < \frac{p_1 + p_2}{1 + p_2}. \tag{16}$$

To better understand this inequality, we plot these bounds on the group credence p^* in Figure (2), below. The plot depicts p^* (z-axis) as a function of p_1 (x-axis) and p_2 (y-axis). We can see that the bounds on p^* are very tight at the extremes, and widest near middling values. This is to be expected because when the valences of the individuals' predictions are near middling values then there are many possible evidence histories consistent with the group's credence – i.e., many different ways the evidence could be overlapping among the group's members. In such cases, our approach is at its most permissible. It allows the group belief to be anywhere between these wide bounds. However, as the individual credences become sharper in their valence

– i.e., closer to 0 or 1 – then our approach narrows down the range of rationally permissible credences the group could adopt.

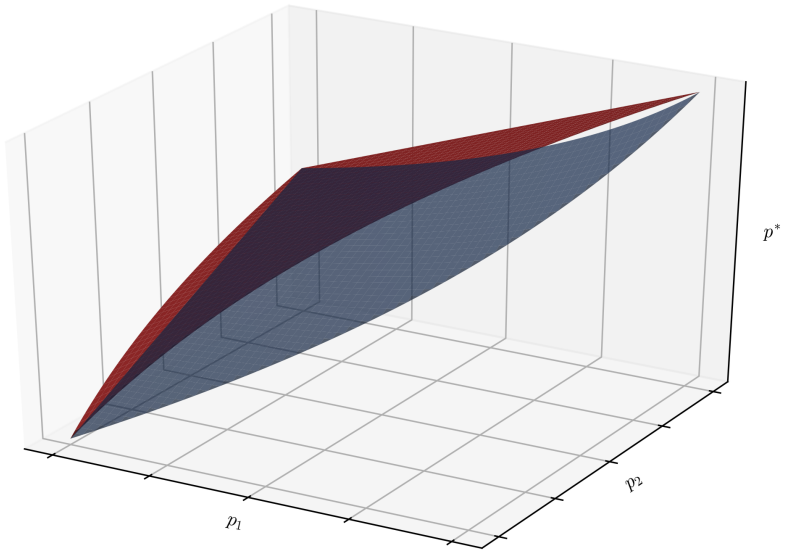

Figure 2: Plot depicting bounds on the group credence, p^*, as a function of A and B's credences, p_1 and p_2, respectively, depending on different estimates about the degree of evidential overlap among the group's members. All axes range from 0 to 1.

We can further illuminate the normative constraints that our approach imposes by considering a few special cases of (16). Consider, first, the case where the individuals' credences are identical ($p_1 = p_2$).

Letting $p_1 = p_2 = p$, (16) reduces to,

$$\frac{p}{2-p} < p^* < \frac{2p}{1+p}. \tag{17}$$

We can now plot these bounds as a two dimensional slice of the above plot, as follows.

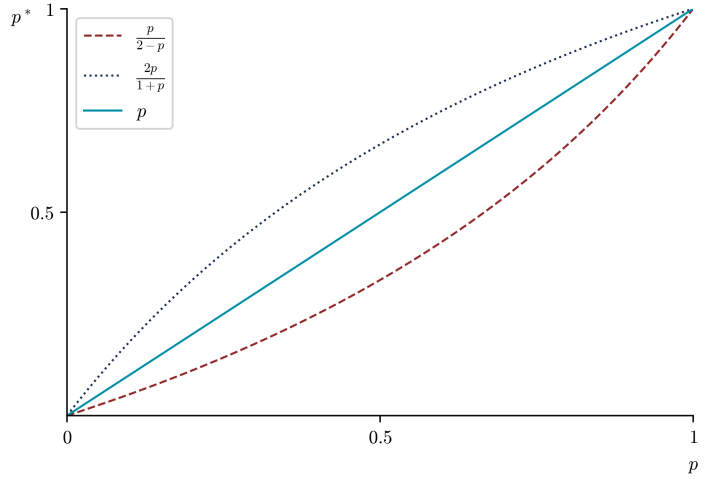

Figure 3: Plot depicting bounds on the group credence, p^*, as a function of A and B's credences, p, depending on different estimates about the degree of evidential overlap among the group's members.

Consider two further cases. When $p_1 = 0$,

$$\frac{1}{2}\,p_2 < p^* < \frac{p_2}{1+p_2}. \tag{18}$$

And when $p_2 = 1$,

$$\frac{1}{2 - p_1} < p^* < \frac{1 + p_1}{2}. \tag{19}$$

By comparison, the simple average is $p_2/2$ when $p_1 = 0$ and $(1 + p_1)/2$ when $p_2 = 1$. Thus, simple averaging here corresponds to only one of many compatible evidence histories.

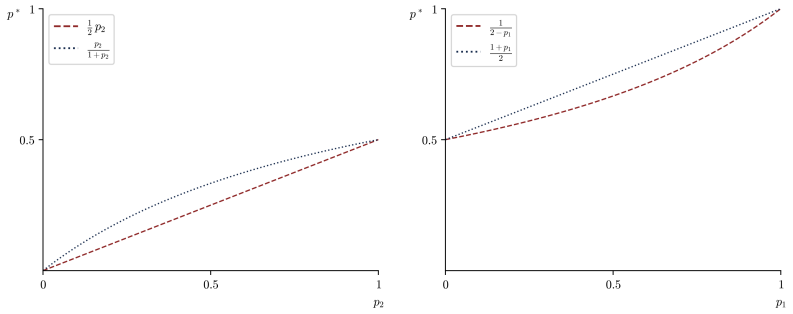

Figure 4: Plot depicting bounds on the group credence, p^*, as a function of A and B's credences, p, depending on different estimates about the degree of evidential overlap among the group's members. In the left panel, $p_1 = 0$, corresponding to (18), and in the right panel, $p_2 = 1$, corresponding to (19).

The point, in short, is that there are two ways to make use of the evidence-first method. The first is where the individuals are able to share their total evidence with each other and discern the degree of its overlap. In such cases, our approach offers essentially a step-by-step recipe for getting to a full group distribution. The general idea here is that we start from the notion that we should use all available information, in the spirit of Good (1967), and we identify a systematic approach to combining that information in a way that is particularly sensitive to avoiding evidential overlap.[18] Moreover, the aggregation method we propose is maximally fine-grained, or informative, in the sense that we identify the full distribution. Using that distribution, the group can then pull out any statistic of interest – such as a mean, a confidence interval, or any quantile.

The second is where the individuals do not know either what the evidence is or the extent of its overlap (or both). In such cases, they can use the individual predictions to construct upper and lower bounds on the rationally permissible group credence. So even though in this case we cannot tell the group where, precisely, to move to, we can tell the group which credences to avoid. This is similar to how Joyce (1998), for example, views the normative role of the accuracy-dominance framework. In that framework, if an agent has incoherent credences, we cannot tell her which coherent credences, precisely, she should adopt. But we can tell her that there are many credences which accuracy-dominate her own, and therefore that she should not remain where she is. Thus, our approach serves as a normative guide in both cases. However, the more information we have about the aggregation problem, the more firmly are the standards of rationality delivered by the evidence-first method.

7. Concluding Remarks

We have presented a general and flexible evidence-driven framework for combining beliefs. The method's core virtues are that the group belief is update commutative and reflects the full range of information available to its individuals while simultaneously taking into account any overlaps in their evidence. Beyond the technical virtues, our hope is to encourage a general rethinking of the belief combination problem, from a question of how to combine numerical credences, to a question of how to identify and appropriately combine evidence.

18. This is particularly important because if the individual members of the group are even slightly correlated, then the incremental value of additional members (i.e., of additional information) fades away rather quickly. For example, see Figure (1) of Clemen and Winkler (1985).

Appendix

Theorem 1 (Update Commutativity). Let $\pi_i(p)$ be i's prior distribution for p, for $i = 1, 2$. Let $\Pi(p)$ be the group prior, derived using (7). Let $\pi_i(p|r, n)$ be i's posterior distribution for p, obtained from $\pi_i(p)$ via Bayes' Rule, after learning new information r and $n - r$, and let $\Pi(p|r, n)$ be the group posterior, obtained from $\Pi(p)$ via Bayes' Rule, also after learning r and $n - r$. Finally, with slight abuse of notation, let $\Pi(p)|r, n$ be the group posterior obtained if we first update $\pi_i(p)$ to $\pi_i(p|r, n)$ and then combine $\pi_i(p|r, n)$ using (7). Then,

$$\Pi(p)|r, n = \Pi(p|r, n). \tag{20}$$

Proof: Suppose, first, we update then combine. We know that each person's priors are given by:

$$\pi_1(p) \sim \text{beta}(\alpha_0, \beta_0),$$
$$\pi_2(p) \sim \text{beta}(\alpha_0, \beta_0).$$

Using Bayes' Rule, we obtain the following individual posteriors:

$$\pi_1(p|r, n) \sim \text{beta}(\alpha_0 + \alpha_1 + \alpha_c, \beta_0 + \beta_1 + \beta_c),$$
$$\pi_2(p|r, n) \sim \text{beta}(\alpha_0 + \alpha_2 + \alpha_c, \beta_0 + \beta_2 + \beta_c).$$

where d is now expressed in terms of α_i, α_c, and β_c. Combining these, we get the following group distribution:

$$\begin{aligned}\Pi(p)|r, n &\sim \text{beta}(\alpha_1 + \alpha_2 + 2(\alpha_c + \alpha_0) - \alpha_c - \alpha_0, \beta_1 + \beta_2 + 2(\beta_c + \beta_0) - \beta_c - \beta_0)\\ &= \text{beta}(\alpha_1 + \alpha_2 + \alpha_c + \alpha_0, \beta_1 + \beta_2 + \beta_c + \beta_0)\\ &= \text{beta}(r^*, n^* - r^*)\\ &= \frac{\Gamma(n^*)}{\Gamma(r^*)\Gamma(n^*)}\ p^{r^*-1}\ (1-p)^{n^*-r^*-1},\end{aligned}$$

where $\Gamma(n) = (n-1)!$.

Suppose, next, we first combine then update, as in Equation 7. We know that each person's priors are given by:

$$\pi_1(p) \sim \text{beta}(\alpha_0, \beta_0),$$
$$\pi_2(p) \sim \text{beta}(\alpha_0, \beta_0).$$

Combining these distributions, we obtain:

$$\begin{aligned}\Pi(p) &\sim \text{beta}(2\alpha_0 - \alpha_0, 2\beta_0 - \beta_0)\\ &\sim \text{beta}(\alpha_0, \beta_0).\end{aligned}$$

If we now update this group distribution, we get the following group posterior:

$$\Pi(p|r, n) \sim \text{beta}(\alpha_0 + \alpha_1 + \alpha_2 + \alpha_c, \beta_0 + \beta_1 + \beta_2 + \beta_c).$$

Note that $\alpha_0 + \alpha_1 + \alpha_2 + \alpha_c = r^*$ and $\beta_0 + \beta_1 + \beta_2 + \beta_c = n^* - r^*$. Hence,

$$\Pi(p|r, n) \sim \text{beta}(r^*, n^* - r^*).$$

As a result,

$$\Pi(p|r, n) = \Pi(p)|r, n.$$

□

Acknowledgments

We would like to thank Frederick Eberhardt, Christopher Hitchcock, Jan-Willem Romeijn and Jonathan Weisberg for their valuable comments, as well as audiences at the UC Irvine Formal Epistemology Workshop, University of Bristol, and the California Institute of Technology.

Funding

Babic's work was supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC), Insight Grant (number 435-2022-0325); a grant from the Government of Hong Kong, SAR China, University Grants Committee, General Research Fund (Project Code 17616324); and a grant from the HKU Musketeers Foundation Institute of Data Science, HKU100 Fund.

References

Aumann, R. J. (1987). Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica 55*(1), 1–18.

Babic, B. (2019). A Theory of Epistemic Risk. *Philosophy of Science 86*(3), 522–550.

Babic, B., A. Gaba, I. Tsetlin, and R. L. Winkler (2024). Noisy Stereotypes. *British Journal for the Philosophy of Science 75*(1), 153–177.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carnap, R. (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Christensen, D. (2011). Disagreement, Question-Begging and Epistemic Self-Criticism. *Philosophers' Imprint 11*(6), 1–22.

Clemen, R. T. (1987). Combining Overlapping Information. *Management Science 33*(3), 373–380.

Clemen, R. T. and R. L. Winkler (1985). Limits for the Precision and Value of Information from Dependent Sources. *Operations Research 2*(33), 427–442.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré 7*(1), 1–68.

Dietrich, F. (2019). A Theory of Bayesian Groups. *Nous 53*(3), 708–736.

Dietrich, F. and C. List (2016). Probabilistic Opinion Pooling. In C. Hitchcock and A. Hájek (Eds.), *Oxford Handbook of Probability and Philosophy*, Chapter 25, pp. 519–542.

Dietrich, F. and K. Spiekermann (2013). Independent Opinions? On the Causal Foundations of Belief Formation and Jury Theorems. *Mind 122*(487), 655–685.

Easwaran, K., L. Fenton-Glynn, C. Hitchcock, and J. D. Velasco (2016). Updating on the Credences of Others: Disagreement, Agreement and Synergy. *Philosophers' Imprint 16*(11), 1–39.

Good, I. (1967). On the Principle of Total Evidence. *The British Journal for the Philosophy of Science 17*(4), 319–321.

Greco, D. and B. Hedden (2016). Uniqueness and Metaepistemology.

Journal of Philosophy 113(8), 365–395.

Hájek, A. (ms). Staying Regular. *Unpublished Manuscript*.

Harsanyi, J. (1987). Games with Incomplete Information Played by Bayesian Players (Parts I, II, III). *Management Science 14*(5), 159–182, 320–334, 486–502.

Hedden, B. (2015). Time-Slice Rationality. *Mind 124*(494), 449–491.

Huttegger, S. M. (2017a). Inductive Learning in Small and Large Worlds. *Philosophy and Phenomenological Research 95*(1), 90–116.

Huttegger, S. M. (2017b). *The Probabilistic Foundations of Rational Learning*. Cambridge: Cambridge University Press.

Jaynes, E. T. (1957a). Information Theory and Statistical Mechanics. I. *Physical Review 106*(4), 620–630.

Jaynes, E. T. (1957b). Information Theory and Statistical Mechanics. II. *Physical Review 108*(2), 171–190.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London: Series A 186*(1007), 453–461.

Johnson, W. (1924). *Logic: Part III, The Logical Foundations of Science*. Cambridge: Cambridge University Press.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science 65*(4), 575–603.

Joyce, J. M. (2005). How Probabilities Reflect Evidence. *Philosophical Perspectives 19*(1), 153–178.

Kinney, D. (2020). Why Average When You Can Stack: Better Methods for Generating Accurate Group Credences. *Manuscript*.

Laplace, P. S. (1814). *Théorie Analytique des Probabilités*. Paris: Courcier.

Lichtendahl, K. C., Y. Grushka-Cockayne, V. R. Jose, and R. L. Winkler (2021). Extremizing and Anti-Extremizing in Bayesian Ensembles of Binary-Event Forecasts. *Operations Research 70*(5), 2597–3033.

Lindley, D. (1983). Reconciliation of Probability Distributions. *Operations Research 31*(5), 866–880.

Lindley, D. V. and L. Phillips (1976). Inference for a Bernoulli Process (A Bayesian View). *The American Statistician 30*(3), 112–119.

List, C. and P. Pettit (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Moss, S. (2011). Scoring Rules and Epistemic Compromise. *Mind 120*(480), 1053–1069.

Moss, S. (2015). Time-Slice Epistemology and Action Under Indeterminacy. In J. Hawthorne and T. S. Gendler (Eds.), *Oxford Studies in Epistemology*, Volume 5, pp. 172–194. Oxford: Oxford University Press.

Pettigrew, R. (2019). On the Accuracy of Group Credences. In T. S. Gendler and J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Chapter 6, pp. 137–160. Oxford: Oxford University Press.

Robert, C. P. (2007). *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation* (2 ed.). New York: Springer.

Romeijn, J. (2024). An Interpretation of Weights in Linear Opinion Pooling. *Episteme 21*(1), 19–33.

Romeijn, J. and D. Atkinson (2011). A Condorcet Jury Theorem for Unknown Jury Competence. *Politics, Philosophy and Economics 10*(3), 237–262.

Russell, J. S., L. Buchak, and J. Hawthrone (2015). Groupthink. *Philosophical Studies 172*(5), 1287–1309.

Skyrms, B. (1980). *Causal Necessity*. London: Yale University Press.

Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy 48*, 1–25.

White, R. (2010). Evidential symmetry and mushy credence. In T. S. Gendler and J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Volume 3, Chapter 7, pp. 161–186. Oxford: Oxford University Press.

Widmer, G. and M. Kubat (1996). Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning 23*(1), 60–101.

Williamson, J. (2010). *In Defense of Objective Bayesianism*. Oxford: Oxford University Press.

Williamson, J. (2019). Aggregating Judgments by Merging Evidence. *Journal of Logic and Computation 19*(3), 461–473.

Zabell, S. L. (2005). *Symmetry and its Discontents: Essays on the History of Inductive Probability*. Cambridge: Cambridge University Press.