Philosophy of Science (2022), **89**, 845–863
doi:10.1017/psa.2022.3

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

ARTICLE

Why Average When You Can Stack? Better Methods for Generating Accurate Group Credences

David Kinney

Princeton University, 222 Peretsman Scully Hall, Princeton, NJ, US
Email: david.kinney@princeton.edu

(Received 18 June 2020; revised 29 June 2021; accepted 28 July 2021; first published online 25 April 2022)

Abstract

Formal and social epistemologists have devoted significant attention to the question of how to aggregate the credences of a group of agents who disagree about the probabilities of events. Moss (2011) and Pettigrew (2019) argue that group credences can be a linear mean of the credences of each individual in the group. By contrast, I argue that if the epistemic value of a credence function is determined solely by its accuracy, then we should, where possible, aggregate the underlying statistical models that individuals use to generate their credence functions, using "stacking" techniques from statistics and machine learning first developed by Wolpert (1992).

1. Introduction

Suppose that Alphonse and Belinda are rushing to catch a train from Brussels to Amsterdam and do not have time to check the schedule. Alphonse's credence that the train leaves before noon is .7, whereas Belinda's credence that the train leaves before noon is .3. As a couple, what is their credence that the train leaves before noon? Formal epistemologists have devoted considerable attention to this type of question, with Lehrer and Wagner (1983), Russell, Hawthorne, and Buchak (2015), and Dietrich and List (2017), among others, producing impossibility results showing that no aggregation rule can satisfy a set of prima facie desirable conditions. Russell, Hawthorne, and Buchak go on to argue that a *geometric* averaging rule can uniquely satisfy an attractive subset of these desiderata. Most importantly, it is shown that geometric averaging rules allow for group credences to commute with

*Many thanks to David Wolpert for first introducing me to the literature on stacking. For their helpful comments on various drafts, I am also grateful to Hein Duijf, Remco Heesen, James Nguyen, Richard Pettigrew, Joe Roussos, Jeremy Strasser, David Watson, Kevin Zollman, and two anonymous reviewers for this journal, as well as audiences at the London School of Economics and Political Science (LSE) Choice Group Seminar, the 2020 Formal Epistemology Workshop, the 2020 Conference on Bayesian Epistemology: Perspectives and Challenges at the Munich Center for Mathematical Philosophy, and the 2020 Workshop on the Wisdom and Madness of Crowds at the Institute for Logic, Language, and Computation at the University of Amsterdam.

© The Author(s), 2022. Published by Cambridge University Press on behalf of the Philosophy of Science Association.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

conditionalization: if each individual in a group updates on the same information and then takes a geometric mean of their posterior credences, their group credence will be the same as it would be if they had first taken the geometric mean of their prior credences and then updated their group credence on the same information. *Linear* methods for credal averaging do not allow for such consistency of conditionalization between individuals and the groups that they comprise, a feature of linear averaging also highlighted by Wagner (1985), Bradley (2007), Jehle and Fitelson (2009), Steele (2012), Staffel (2015), and Kuan (2021). By contrast, Moss (2011) and Pettigrew (2019) argue for the alternative thesis that linear averaging is the superior method for aggregating group credences. This thesis is defended in spite of the problems with conditionalization just mentioned.

My aim in this article is not to weigh in on either side of the debate between those who believe that group credences should be generated via geometric averaging of individual credences and those who believe that group credences should be generated via linear averaging of individual credences. Simply put, I argue that both camps are in the wrong, at least insofar as they aim to provide general normative guidelines for generating group credences. My argument for this claim proceeds as follows: Like Pettigrew (2019), I hold that when devising a method to generate group credences, we ought to favor methods that allow the group to be as accurate as possible, at least insofar as we care about the epistemic value of the group's credences. However, I show that there are cases in which no averaging method can produce very accurate group credences. Informally, these are cases in which all individuals in a group assign credences to events based on inaccurate models of the relevant data. One might think that in such cases, no method of arriving at group credences can be expected to be accurate, but recent work on model stacking by Le and Clarke (2017) shows that this is not the case. Although Le and Clarke do not specifically discuss the aggregation of credences, I apply their techniques to show how, even when all members of a group use inaccurate models to assign credences to events, the group can use stacking techniques to arrive at credences that are more accurate than those that would be generated by any averaging method.

Here is the plan for this article: In section 2, I provide the formal background needed to make my argument. In section 3, I demonstrate that when all individuals in a group have inaccurate models of a given data-generating process, neither geometric nor linear averaging of credences produces accurate group credences. In section 4, I show how stacking techniques allow for a group to generate more accurate credences by aggregating the statistical models from which each individual derives their credences, even when each of these models is individually highly inaccurate. I therefore conclude that group credences should, where possible, be generated by stacking individuals' models rather than averaging individuals' credences. In section 5, I address Pettigrew's unanimity condition on credal aggregation, which is central to his argument that groups that care only about being accurate should pool their credences via linear averaging. I argue that there is no purely accuracy-based reason to accept unanimity as a constraint on credal aggregation methods. I also address similar arguments due to Moss (2011). In section 6, I respond to possible counterarguments to my proposal; in so doing, I highlight the value of asking individuals for the *reasons* behind their partial beliefs when attempting to determine a group credence. In section 7, I offer concluding remarks.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

2. Formal preliminaries

2.1 Group credences as means

Throughout this article, I will represent the problem of determining group credences from individual credences as follows: A group of individuals $I = \{1, 2, \ldots, N\}$ shares a common sample space Ω, or set of possible worlds, and shares a common algebra $\mathcal{A}_\Omega$ on Ω, or a collection of subsets of Ω that is closed under complement, union, and intersection. Each individual $i \in I$ plans to have their own credence function $Cr_i : \mathcal{A}_\Omega \to [0, 1]$ that obeys the standard Kolmogorov axioms. Thus, each individual has their own credal probability space $\mathbf{Cr}_i = (\Omega, \mathcal{A}_\Omega, Cr_i)$. For a sample space Ω, an algebra $\mathcal{A}_\Omega$, a set of individuals I, and a partition $\mathcal{F}$ of Ω such that for each $F \in \mathcal{F}$, $F \in \mathcal{A}_\Omega$, the *group credence problem* $(\Omega, \mathcal{A}_\Omega, I, \mathcal{F})$ is to find a single credence function Cr^* that represents the credence of the entire group I in each element of $\mathcal{F}$.

To illustrate using the previous example, we will begin with a set of two individuals $I = \{\text{Alphonse, Belinda}\}$, each of whom shares a sample space Ω and algebra $\mathcal{A}_\Omega$. We define a partition $\mathcal{F} = \{F, \neg F\}$, where F is the set of worlds in Ω in which the train from Brussels to Amsterdam leaves before noon, and $\neg F$ is its complement in Ω (i.e., the set of worlds in which the train leaves at noon or later). We know that Alphonse's credence function Cr_A is such that $Cr_A(F) = .7$ and $Cr_A(\neg F) = .3$, and Belinda's credence function Cr_B is such that $Cr_B(F) = .3$ and $Cr_B(\neg F) = .7$. The problem of finding Alphonse and Belinda's group credence in each element of the partition $\{F, \neg F\}$ is represented as $(\Omega, \mathcal{A}_\Omega, I, \mathcal{F})$.

In formal epistemology, it is typically assumed that the solution to the group credence problem is to calculate a *mean credence* for each element of the relevant partition. For any element F of a partition $\mathcal{F}$ of the relevant sample space and any set of credences $\{Cr_1(F), \ldots, Cr_N(F)\}$, a mean $\mu(\{Cr_1(F), \ldots, Cr_N(F)\})$ of those credences satisfies the following two individually necessary and jointly sufficient conditions:

Homogeneity: $\mu(\{tCr_1(F), \ldots, tCr_N(F)\}) = t^\alpha \mu(\{Cr_1(F), \ldots, Cr_N(F)\})$, for any $t \in \mathbb{R}$ and some $\alpha \in \mathbb{R}$.

Min-max: $\min\{Cr_1(F), \ldots, Cr_N(F)\} \leq \mu(\{Cr_1(F), \ldots, Cr_N(F)\}) \leq \max\{Cr_1(F), \ldots, Cr_N(F)\}$.

The two primary types of means considered in the formal epistemology literature on group credences are the *linear* and *geometric* means. They are defined as follows:

Linear mean: $\mu_L(\{Cr_1(F), \ldots, Cr_N(F)\}) = \sum_{i=1}^{N} w_i Cr_i(F)$, where $\{w_1, \ldots, w_N\}$ is a set of nonnegative, individual-specific weights such that $\sum_{i=1}^{v} w_i = 1$.

Geometric mean: $\mu_G(\{Cr_1(F), \ldots, Cr_N(F)\}) = (\sqrt[k]{\prod_{i=1}^{N} Cr_i(F)^{w_i}}) / (\sum_{F \in \mathcal{F}} \sqrt[k]{\prod_{i=1}^{N} Cr_i(F)^{w_i}})$, where $\{w_1, \ldots, w_N\}$ is a set of nonnegative, individual-specific weights such that $\sum_{i=1}^{v} w_i = k$.

For a set of individuals I and a partition $\mathcal{F}$ of their shared sample space, if a group credence function Cr^* is such that for each $F \in \mathcal{F}$, $Cr^*(F) = \mu_L(\{Cr_1(F), \ldots, Cr_N(F)\})$, then Cr^* necessarily obeys the Kolmogorov axioms, provided that the same set of individual-specific weights $\{w_1, w_2, \ldots, w_N\}$ is used to calculate each group credence

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

$Cr^*(F)$. Similarly, if a group credence function Cr^* is such that for each $F \in \mathcal{F}$, $Cr^*(F) = \mu_G(\{Cr_1(F), \ldots, Cr_N(F)\})$, then Cr^* necessarily obeys the Kolmogorov axioms, provided that the same set of individual-specific weights $\{w_1, w_2, \ldots, w_N\}$ is used to calculate each group credence $Cr^*(F)$ and that there is an event $F \in \mathcal{F}$ that has nonzero probability according to each individual's credence function (Dietrich 2019).

As mentioned in the introduction, my goal in this article is not to adjudicate which of these two means provides the better mechanism for credal aggregation. Rather, my focus will be on what these two means have in common, with a particular focus on the min-max condition that both means satisfy. In what follows, I will argue that the min-max condition constrains the possible accuracy of a group credence in cases where all members of the group have highly inaccurate credences. I will show that this is not the case for stacking-based methods of determining group credences.

Throughout this article, I assume that all individuals in the group credence problem have a shared data set. This assumption means that one must be precise about when in the belief-formation and data-collection process the group members must make their decision about how they will solve the group credence problem. Both Kadane and Lichtenstein (1982) and Dawid (1982) prove results showing that if

1. an agent's beliefs are represented by a probability space $(\Omega, \mathcal{A}, Cr)$;
2. $(\theta_1 \ldots, \theta_n)$ is a sequence of random variables with the same range that are measurable with respect to $(\Omega, \mathcal{A}, Cr)$;
3. the agent sets their credence in $\theta_{i+1} = x$, where $i + 1 \leq n$, for any x in the range of θ_{i+1}, by conditionalizing on the outcome of all previous variables $(\theta_1, \ldots, \theta_i)$;
4. π_x is the proportion of random variables in $(\theta_1 \ldots, \theta_n)$ with outcome x; and
5. ζ is a random variable measurable with respect to $(\Omega, \mathcal{A}, Cr)$ such that $\zeta = \lim_{n \to \infty} [\pi_x - 1n \sum_{i=1}^{n} Cr(\theta_i = x)]$,

then $Cr(\zeta = 0) = 1$. Thus, Bayesian agents take themselves to be maximally well calibrated with a given data source, such that in the long run, their credence that a given observation will return a given result will converge to the proportion of cases in which that output does indeed have that result. Thus, they will not update their credences at all on the basis of *another* agent's credences when they know that the second agent has access to all and only the same data sources that they do. Note that this applies even when $i = 0$, or when no agents have observed any data, but they have formed prior credences over the set of possible data streams that they might observe.

Thus, we must be careful about *when* a group adopts its solution to the group credence problem. That is, we assume throughout that a group must agree on its solution to the group credence problem from an ex ante position—before any data have been collected and before any individual prior credences have been adopted by the individuals in the group. This is why, in the introduction to the group credence problem, I say that each individual in the group *plans* to have their own credence function Cr_i over the shared algebra $\mathcal{A}_\Omega$. They then agree to be bound by their planned solution to the group credence problem at some future point in time, even though, at that point, each group member i will regard this solution as suboptimal unless it returns precisely their individual credence Cr_i. Note that this applies equally to solutions to the group credence problem that take a mean of individual credences and the

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

stacking-based solution that I defend later in the article in cases where individuals share all of their data.

2.2 Inaccuracy

My argument depends on our ability to make comparisons between group credence functions produced by different methods with respect to their (in)accuracy. Thus, I need to introduce the formal notion of an *inaccuracy measure*. Generally, an inaccuracy measure $\mathcal{I}$ is a real-valued function that takes as its arguments a credence function Cr; a partition $\mathcal{F}$ of the sample space Ω, where Cr is defined over an algebra $\mathcal{A}_\Omega$ on Ω; and an event $F^\dagger \in \mathcal{F}$. We interpret $F^\dagger$ as the element of the partition $\mathcal{F}$ that contains the actual world. The higher the value of $\mathcal{I}(Cr, \mathcal{F}, F^\dagger)$, the more inaccurate Cr is with respect to the probabilities that it assigns to the elements of $\mathcal{F}$. The two inaccuracy measures discussed most in formal epistemology are the *Brier measure* and the *logarithmic measure*. Let $\mathcal{F} = \{F_1, \ldots, F_m\}$, and let $T : \mathcal{F} \rightarrow \{0, 1\}$ be a truth value function such that $T(F_j)$=1 if F_j=$F^\dagger$, and $T(F_j)$=0 otherwise. The Brier measure and logarithmic measure are defined as follows:

Brier measure: $\mathcal{I}_B(Cr, \mathcal{F}, F^\dagger) = \frac{1}{m}\sum_{j=1}^{m} (Cr(F_j) - T(F_j))^2$.
Logarithmic measure: $\mathcal{I}_L(Cr, \mathcal{F}, F^\dagger) = -\ln(Cr(F^\dagger))$.

Although there are formal differences between the two scores, these differences are not directly relevant to my argument. As such, I will use both measures when assessing the accuracy of any given group credence.

In addition to measuring the accuracy of a credence function when the actual world is in a particular element of a partition $\mathcal{F}$, we can also calculate the expected inaccuracy of any credence function Cr, where the expectation is calculated according to some probability distribution P that is defined over the same algebra $\mathcal{A}_\Omega$ and the sample space Ω as Cr. This expectation is defined via the following equation:

Expected inaccuracy: $\mathbb{E}_P(\mathcal{I}(Cr, \mathcal{F}, \cdot)) = \sum_{j=1}^{m} P(F_j)\mathcal{I}(Cr, \mathcal{F}, F_j)$.

This definition assumes that we interpret a probability $P(F_j)$ as expressing the probability that F_j contains the actual world. Note that we use a placeholder instead of $F^\dagger$ when calculating the expected inaccuracy because calculating an expected accuracy assumes that we do not know which element of $\mathcal{F}$ contains the actual world. Because expected inaccuracy is itself a linear mean of inaccuracy measures, it satisfies the min-max condition defined previously.

3. When a mean is not accurate

To show how assigning group credences using a mean can result in highly inaccurate group credences, consider the following case: turning to a new example, let us suppose that Alphonse and Belinda are each reading reports showing the number of lynxes and the number of hares in a given area of forest on a given day. Let us represent each day's report as a pair (l, h), where l is the number of lynxes, and h is the number of hares. The first 4 days produce a data set $\mathcal{D}$ of reports such that $\mathcal{D} = \{(1, 5), (2, 11), (3, 17), (4, 20)\}$. Alphonse and Belinda are then told that on the

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

fifth day of observation, there were five lynxes in the area, but the number of hares could not be measured. They are asked to assign a credence to the event that the number of observed hares was less than 10 and to the event that the number of observed hares was greater than or equal to 10. Each agent completes their task by constructing their own signal–noise model of the data. Their respective models are defined as follows, where the noise term ε is normally distributed around 0 with a standard deviation of $\sigma = 4$:[1]

$$M_A : h = 2l + \epsilon. \tag{1}$$

$$M_B : h = 3l + \epsilon. \tag{2}$$

Thus, when $l = 5$, Alphonse and Belinda's estimates of the number of hares in the area will be normally distributed, with a standard deviation of $\sigma = 4$, around the mean of $2(5) = 10$ and $3(5) = 15$, respectively.

Alphonse and Belinda share a sample space Ω that contains the positive integers, representing the number of hares observed on the fifth day. Their credence functions are defined over an algebra $\mathcal{A}_\Omega$, which we take to be the power set of Ω. We partition Ω into the set $\mathcal{H} = \{H_{<10}, H_{\geq 10}\}$, where $H_{<10}$ is the set of worlds in which there are less than 10 hares in the area, and $H_{\geq 10}$ is the set of worlds in which there are 10 or more hares in the area. In keeping with the models specified previously, Alphonse and Belinda's credences in each of the two events in this partition can be calculated as follows:[2,3]

$$Cr_A(H_{<10}) = \int_0^{9.5} \frac{1}{4\sqrt{2\pi}} e^{-(h-10)^2/2(4^2)} \mathrm{dh} \approx .44. \tag{3}$$

$$Cr_A(H_{\geq 10}) = 1 - Cr_A(H_{<10}) \approx .56. \tag{4}$$

$$Cr_B(H_{<10}) = \int_0^{9.5} \frac{1}{4\sqrt{2\pi}} e^{-(h-15)^2/2(4^2)} \mathrm{dh} \approx .08. \tag{5}$$

$$Cr_B(H_{\geq 10}) = 1 - Cr_B(H_{<10}) \approx .92. \tag{6}$$

Thus, if Alphonse and Belinda's joint credence function Cr^* is generated by taking a mean of their individual credences, then because of the min-max constraint on a mean, it must be the case that $.08 \leq Cr^*(H_{<10}) \leq .44$ and $.56 \leq Cr^*(H_{\geq 10}) \leq .92$.

[1] Throughout this article, I assume that all error terms have the standard deviation $\sigma = 4$. This assumption is for mathematical tractability, and it is not a requirement for my argument that all signal–noise models have an error term with the same standard deviation.

[2] To briefly explain these calculations, note that the normal distribution $f(x)$ has the form $f(x) = (1/(\sigma\sqrt{2\pi}))e^{-(x-\mu)^2/2\sigma^2}$, where σ is the standard deviation of the distribution, and μ is the mean. The probability that $f(x) \in [y, z]$ can be calculated by taking the integral $\int_y^z f(x)\mathrm{dx}$. I estimate the probability that fewer than 10 hares are observed by taking the integral $\int_0^{9.5} f(h)\mathrm{dh}$. The use of normally distributed, real-valued error terms is an idealization because it allows for noninteger and negative values of hares, although states in which the number of hares is negative have a very low probability according to all of the models considered here.

[3] All calculations performed in this article are reproduced in the Jupyter notebook available at https://github.com/davidbkinney/whyaveragewhenyoucanstack.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

Table 1. Maximum and minimum expected inaccuracy of Alphonse and Belinda's mean credence under both measures, as compared to expected inaccuracy of the true probability distribution, according to itself.

	Brier	Logarithmic
$\max \mathbb{E}_{P_T}(\mathcal{I}(Cr^*_M, \mathcal{H}, \cdot))$	.394	.587
$\min \mathbb{E}_{P_T}(\mathcal{I}(Cr^*_M, \mathcal{H}, \cdot))$	.014	.088
$\mathbb{E}_{P_T}(\mathcal{I}(P_T, \mathcal{H}, \cdot))$	1.06619×10^{-4}	5.78×10^{-4}

However, Alphonse and Belinda are both poor data analysts. By stipulation, the true data-generating process is modeled as follows, where ε is also normally distributed around 0 with a standard deviation of $\sigma = 4$:

$$M_T : h = 5l + \epsilon. \tag{7}$$

This means that the true probability distribution P_T over $\mathcal{H}$ can be calculated as follows:

$$P_T(H_{<10}) = \int_0^{9.5} \frac{1}{4\sqrt{2\pi}} e^{-(h-25)^2/2(4^2)} \mathrm{dh} \approx 5.33 \times 10^{-5}. \tag{8}$$

$$P_T(H_{\geq 10}) = 1 - P_T(H_{<10}) \approx .99995. \tag{9}$$

Table 1 shows the maximum and minimum expected inaccuracy of Alphonse and Belinda's group credence Cr^* according to the true probability distribution P_T, for both the Brier and logarithmic inaccuracy scores, under the assumption that their group credence must be a mean of their individual credences. It also shows the expected inaccuracy of the true probability distribution, according to itself. It should be clear from the table that even in the best-case scenario, Alphonse and Belinda's mean credence will be quite inaccurate compared to the true probability distribution.

One might be tempted to think that this limit on the expected accuracy of Alphonse and Belinda's joint credence is an inevitable consequence of their being poor data analysts. If all individuals in a group are in a poor epistemic state, it could be argued, why should we expect the group as a whole to fare well? In the next section, I will show how stacking techniques for generating group credences demonstrate that group inaccuracy is not an inevitable consequence of unanimous individual inaccuracy.

4 How stacking leads to more accurate group credences

Let us once again reference the lynx-and-hare example from the previous section, wherein Alphonse and Belinda's models of the observed data are given by M_A and M_B. We want to generate a "stacked" model M_S, which we hope will yield more accurate credences than either of the individual models. Our strategy will be to come up with a vector of stacking weights $\vec{\mathfrak{w}} = \{\mathfrak{w}_1, \mathfrak{w}_2\}$ such that the group will make predictions using the following stacked model:

$$M_S : h = \mathfrak{w}_1 2l + \mathfrak{w}_2 3l + \epsilon. \tag{10}$$

Note that ϵ is once again an error term normally distributed around 0 with a standard deviation of $\sigma = 4$. Importantly, these weights are not required to sum to 1; see Clyde

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

and Iversen (2013, 485–86) for a simple demonstration of cases in which the sum-to-1 constraint limits the accuracy of a stacked model. Thus, there is a large space of weights to choose from, and we will need to choose carefully in order to arrive at a model that generates more accurate credences than any mean.

In order to do this, we will need to add an additional piece to both Alphonse's and Belinda's modeling repertoires. Let $\mathcal{D}$ be the set of all possible data sets $D = \{(x_1, y_1), \ldots, (x_v, y_v)\}$ that an agent might observe, and let $\mathcal{L}_i : \mathcal{D} \rightarrow \mathcal{M}$ be a given individual i's *learning algorithm*, where $\mathcal{M}$ is the set of all possible signal–noise models of the form $f(x) + \varepsilon$. Individuals use their learning algorithms to build models of the data-generating process from any given data set. Let $D_{-\alpha}$ denote the data set $D \backslash \{(x_\alpha, y_\alpha)\}$ (i.e., the data set D with the αth data point removed). For any data set D, individual i, and data point (x_α, y_α), let $\mathcal{L}_i(D_{-\alpha}) = f_{D,i}^{-\alpha}(x) + \varepsilon$. For any individual i and data set $D = \{(x_1, y_1), \ldots, (x_v, y_v)\}$, their leave-$\alpha$-out vector $\vec{z}_i$ is defined as follows: $\vec{z}_i = [f_{D,i}^{-1}(x_1), \ldots, f_{D,i}^{-v}(x_v)]^{\mathrm{T}}$.

Le and Clarke (2017, 817) prove that with respect to the goal of accurately predicting the value of y_{v+1}, given the value of x_{v+1}, the optimal stacking-weight vector for N individuals and data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_v, y_v)\}$ can be derived as follows: Let $\mathbf{Q}$ be an $N \times N$ matrix such that each entry q_{kl} is given by the following formula:

$$q_{kl} = \sum_{\alpha=1}^{v} f_{D,k}^{-\alpha}(x_\alpha) f_{D,l}^{-\alpha}(x_\alpha). \tag{11}$$

That is, $\mathbf{Q}$ is a matrix such that the entry in the kth row and the lth column is the dot product of the leave-α-out vectors for the individuals k and l. Next, let $\vec{c}$ be the following vector:

$$\vec{c} = \left[\sum_{\alpha=1}^{v} y_\alpha f_{D,1}^{-\alpha}(x_\alpha), \ldots, \sum_{i=1}^{v} y_\alpha f_{D,N}^{-\alpha}(x_\alpha) \right]^{\mathrm{T}}. \tag{12}$$

In other words, $\vec{c}$ is a vector such that each entry is the dot product of a vector containing the actual value of y for each entry in the data set and the leave-α-out vector for each individual in the set. A set of weights that will yield a highly accurate stacking model for the ensemble of models can be found via the following equation:

$$\vec{\mathfrak{w}} = \mathbf{Q}^{-1}\vec{c}. \tag{13}$$

Thus, the optimal stacking weights $\vec{\mathfrak{w}}$ are found by multiplying the inverse of $\mathbf{Q}$ by $\vec{c}$. More precisely, Le and Clarke (2017) prove the following:

Proposition 1. (Le and Clarke 2017, 817). *For any data set* $D = \{(x_1, y_1), \ldots, (x_v, y_v)\}$ *and set of individuals* I, *the weight vector* $\vec{\mathfrak{w}}$ *that uniquely minimizes* $\sum_{\alpha=1}^{v}(y_\alpha - \sum_{i=1}^{N} \mathfrak{w}_i f_{D,i}^{-\alpha}(x_\alpha))^2$ *is* $\vec{\mathfrak{w}} = \mathbf{Q}^{-1}\vec{c}$.

Thus, a learning algorithm $\mathcal{L}_S$ that uses the leave-α-out vector produced by each individual's algorithm to produce a stacked predictive model with weights $\vec{\mathfrak{w}} = \mathbf{Q}^{-1}\vec{c}$ performs optimally well at minimizing the sum of the squared difference between each predicted outcome $f_{D,i}^{-\alpha}(x_\alpha)$ and each actual outcome y_α when the data

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

point (x_α, y_α) is left out of the data set, across all such leave-α-out scenarios. If the probability distribution $\mathbb{P}$ over all possible data sets $D = \{(x_1, y_1), \ldots, (x_v, y_v)\}$ is such that all finite permutations of any data set have equal probability (i.e., all observation sets are "exchangeable"), then by the law of large numbers (see Bernardo and Smith 1994, 403–4), the following also holds:

$$\mathbb{P}(\lim_{\mathrm{v}\to\infty} \frac{1}{\mathrm{v}} \sum_{\alpha=1}^{\mathrm{v}} (\mathrm{y}_\alpha - \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathfrak{w}_\mathrm{i} \mathrm{f}_{\mathrm{D,i}}^{-\alpha}(\mathrm{x}_\alpha))^2 = 0) = 1. \tag{14}$$

Thus, in the infinite limit, we should expect the learning algorithm $\mathcal{L}_S$, which generates signal–noise models by aggregating the existing signal–noise models in the ensemble using the weight vector $\vec{\mathfrak{w}} = \mathbf{Q}^{-1}\vec{c}$, to yield a more accurate prediction $f_{D,S}^{-(v+1)}(x_{v+1})$ of the value y_{v+1}, given the data set $D = \{(x_1, y_1), \ldots, (x_v, y_v)\}$, than any other aggregation of each individual's models.

Applying this method to our running case study, let us suppose that Alphonse and Belinda's common data set is still $D = \{(1, 5), (2, 11), (3, 17), (4, 20)\}$ and that their leave-α-out vectors are specified as follows:

$$\vec{z}_A = [3, 4, 6, 7]^{\mathrm{T}}. \tag{15}$$

$$\vec{z}_B = [4, 5, 9, 11]^{\mathrm{T}}. \tag{16}$$

We use equation (13) to derive the stacking-weight vector $\vec{\mathfrak{w}} = [.74, 1.35]^{\mathrm{T}}$. Thus, the stacked model has the following form:

$$M_S : h = (.74)2l + (1.35)3l + \varepsilon = 5.53l + \varepsilon. \tag{17}$$

This means that when $l = 5$, a credence function over possible values of h that is consistent with the stacked model M_S will be derived from a normal distribution around a mean of $(5.53)5 = 27.65$. Suppose again that this normal distribution has a standard deviation of $\sigma = 4$. If we use the stacked model M_S to derive a joint credence for Alphonse and Belinda over the partition $\mathcal{H} = \{H_{<10}, H_{\geq 10}\}$, we obtain the following:

$$Cr_S^*(H_{<10}) = \int_0^{9.5} \frac{1}{4\sqrt{2\pi}} e^{-(h-27.65)^2/2(4^2)} \mathrm{dh} \approx 2.85 \times 10^{-6}. \tag{18}$$

$$Cr_S^*(H_{\geq 10}) = 1 - P_T(H_{<10}) \approx .999997. \tag{19}$$

The expected inaccuracies of this stacking-derived group credence Cr_S^*, with respect to the true probability distribution P_T, for the Brier and logarithmic measures, are given by the following equations:

$$\mathbb{E}_{P_T}(\mathcal{I}_{\mathcal{B}}(Cr_S^*, \mathcal{H}, \cdot)) = 1.06624 \times 10^{-4}. \tag{20}$$

$$\mathbb{E}_{P_T}(\mathcal{I}_{\mathcal{L}}(Cr_S^*, \mathcal{H}, \cdot)) = 6.84 \times 10^{-4}. \tag{21}$$

Thus, the expected inaccuracy of the stacking-derived joint credence for Alphonse and Belinda, according to the true probability distribution with respect to the number of hares in the region when five lynxes are observed, is much lower than the best-case scenario for any mean-derived joint credence, regardless of whether the Brier measure or the logarithmic measure is used to measure accuracy. Indeed, when the Brier measure is used, the expected inaccuracy of the stacking-derived joint credence function according to the true probability distribution is very close to the expected

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

inaccuracy of the true probability distribution according to itself (recall that the latter expectation is $\mathbb{E}_{P_T}(\mathcal{I}_B(P_T, \mathcal{H}, \cdot)) = 1.06619 \times 10^{-4}$).

As discussed in the introduction, I suppose here that the sole normative expectation for any solution to the group credence problem is that the solution ought to be as accurate as possible. Under this supposition, the preceding example shows that it is at least sometimes the case that the group should aggregate its credences by stacking rather than credal averaging. Notably, this is a case in which we aggregate the values of a random variable predicted by the individuals in a group, where each individual predicts the value of the variable by sampling from their particular credal probability distribution. A group credence is then derived from the aggregate model predicting the value of that random variable. The resulting group credal distribution is, in statistical terminology, a "convolution" of each individual credence function. By contrast, solutions to the group credence problem that take a mean of individual credences produce a group credence that is a "mixture" of individual credence functions. The preceding example shows that when all members of a group are inaccurate, a convolution of group credences can be preferable to a mixture for groups that care about the accuracy of their group credence function. Where some individuals in the group are more accurate, it may be that a mixture of credences is preferable as a solution to the group credence problem. However, in these cases, it is also possible that both convolution-based and mixture-based methods will lead to similar results. For my purposes here, the upshot of this section is that there are at least some cases where a convolution of individual credences, rather than a mixture, is preferable as a solution to the group credence problem.

5. Against unanimity

In the introduction, I mention that Pettigrew (2019) argues in favor of the claim that if the sole goal of a group of agents is to have as accurate of a group credence as possible, then they ought to plan to solve the group credence problem by linear averaging. However, the results in the previous section seem to indicate that this is not the case; when all members of a group have inaccurate credences based on flawed models, stacking results in more accurate credences than linear averaging. So what has gone wrong? In this section, I argue that a premise of Pettigrew's argument—namely, that group credences must satisfy a *unanimity* constraint—need not be satisfied by groups of agents that take accuracy to be the only valuable property of their group credence function.

Let us reconstruct Pettigrew's (2019) argument. His first premise is that any measure of inaccuracy must be a sum of the values of a *strictly proper* and *continuous* scoring rule for each element in a partition under a given probability distribution. A scoring rule is a function $\mathfrak{s}$ such that, for a given element F of a partition $\mathcal{F}$ and a given credence function Cr, $\mathfrak{s}$ takes as its input the truth value $T(F)$ and the credence $Cr(F)$ and returns a value $\mathfrak{s}(T(F), Cr(F))$ between 0 and 1. Recall that $T(F) = 1$ if the actual world is in F, and $T(F) = 0$ otherwise. *Strict propriety* is defined formally as follows:

> **Strict propriety:** For any element F of any partition $\mathcal{F}$ and any two credence functions Cr and Cr', $Cr(F)\mathfrak{s}(1, Cr'(F)) + (1 - Cr(F))\mathfrak{s}(0, Cr'(F))$ is uniquely minimized when $Cr(F) = Cr'(F)$.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

Table 2. Expected inaccuracy, by Alphonse and Belinda's lights, of the maximally accurate group credence that can be derived by taking a linear mean (Cr^*_M) and the stacking-derived group credence (Cr^*_S).

	Brier	Logarithmic
$\mathbb{E}_{Cr_A}\mathcal{I}(Cr^*_M, \mathcal{H}, \cdot)$	.752	1.15
$\mathbb{E}_{Cr_A}\mathcal{I}(Cr^*_S, \mathcal{H}, \cdot)$	.888	5.67
$\mathbb{E}_{Cr_B}\mathcal{I}(Cr^*_M, \mathcal{H}, \cdot)$	.155	.290
$\mathbb{E}_{Cr_B}\mathcal{I}(Cr^*_S, \mathcal{H}, \cdot)$	.169	1.08

> In other words, a scoring rule is strictly proper if, once an agent adopts a certain credence that the actual world is in a given element F of a partition, that agent cannot achieve a lower expected value for that scoring rule by changing their credence in F.

Continuity is defined formally as follows:

> **Continuity:** For any element F of any partition $\mathcal{F}$ and any credence function Cr, $\mathfrak{s}(1, Cr(F))$ and $\mathfrak{s}(0, Cr(F))$ are both continuous functions of $Cr(F)$.

I take this formal definition to be self-explanatory. Because both the Brier and logarithmic scoring rules are sums of the values of a strictly proper and continuous scoring rule for each element in a partition under a given probability distribution, my argument in favor of stacking is consistent with Pettigrew's (2019) first premise, and indeed, nothing that I say here should be taken to dispute it.

Instead, I take issue with Pettigrew's (2019) second premise, which is that group credences must respect *unanimity*. Pettigrew's unanimity constraint can be defined formally as follows:

> **Unanimity:** For any group credence problem $(\Omega, \mathcal{A}_\Omega, I, \mathcal{F})$ and any two credence functions Cr^* and Cr', if for all $i \in I$, $\mathbb{E}_{Cr_i}(\mathcal{I}(Cr^*, \mathcal{F}, \cdot)) < \mathbb{E}_{Cr_i}(\mathcal{I}(Cr', \mathcal{F}, \cdot))$, then Cr' cannot be the group credence function.

This definition formalizes Pettigrew's definition of unanimity as the premise that "if, by the lights of every individual in a group, the expected epistemic value of one credence function is higher than the expected epistemic value of another credence function, then the latter cannot be the credence function of that group" (2019, 145). Pettigrew then proves that if inaccuracy is measured by taking a sum of strictly proper and continuous scoring rules, then any solution Cr' to the group credence problem that is not a linear mean of individual credences will violate unanimity because there will always exist an alternative group credence Cr^* that is (1) a linear mean of the individual credences and (2) unanimously expected to be more accurate than Cr'. Further, any solution to the group credence problem that is a linear mean of individual credences will not be disqualified by the unanimity constraint. Pettigrew

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

takes this to show that agents ought to solve the group credence problem by taking a linear mean of their individual credences.

In the example discussed in the previous section, the group credence arrived at by stacking does not satisfy unanimity. Table 2 shows the expected inaccuracy, by both Alphonse and Belinda's lights, of the maximally accurate group credence Cr^*_M that can be derived by taking a linear mean of their individual credences, as compared to the expected inaccuracy, by both Alphonse and Belinda's lights, of the stacking-derived group credence Cr^*_S. Clearly, both Alphonse and Belinda expect Cr^*_M to be the more accurate credence, so unanimity would rule out Cr^*_S as a solution to their group credence problem. Nevertheless, I have shown in the previous section that the stacking-derived group credence has far greater expected accuracy, by the lights of the true probability distribution over $\mathcal{H}$, than the maximally accurate credence that can be derived by taking a linear mean of Alphonse and Belinda's credences. Thus, it follows that if one believes, as Joyce (1998), Goldman (2001), and Pettigrew (2019) all do, that accuracy is the sole source of value for a credence function, then groups ought to abandon unanimity as an ex ante constraint on how they ought to plan to solve the group credence problem in order to allow for more accurate, stacking-based solutions.

There may be other reasons why groups should plan to solve the group credence problem using a method that is consistent with unanimity. Unanimity might be a necessary condition for finding a solution that reflects a commitment to *both* the claim that the epistemic value of a credence function is its accuracy *and* the claim that group credences should be arrived at through a procedure of deliberative democracy. To accept this, one would have to grant that if all individuals in a group care about accuracy, and all of those individuals expect that a certain credence will be inaccurate, then the group ought to avoid adopting that credence. This conditional may be true, but note that its normative consequent cannot follow from its descriptive antecedent just because the group treats accuracy as the sole virtue of a credence function. This much is shown by the example in the previous section. However, such a normative implication could be valid if the group has a background commitment to the idea that group beliefs should, at a minimum, cohere with the group consensus. Note that accuracy plays no role in this background commitment, which is only about procedural norms. Thus, Pettigrew (2019) is able to give an accuracy-based argument for the claim that group credences should be derived by taking a linear mean of individual credences only by adopting an assumption that is not motivated by concerns relating to the accuracy of credence functions.

Moss (2011) offers a different accuracy-based argument for solving the group credence problem by taking a linear average of individual credences. Her argument proceeds as follows: Suppose that the inaccuracy measure $\mathcal{I}$ is a sum of strictly proper scoring rules. Suppose further that for any partition $\mathcal{F}$ and any group of individuals $I = \{1, \ldots, N\}$, the expected inaccuracy of the group credence Cr^*, according to Cr^*, is given by the following linear mean of the expected inaccuracy of individuals in the group:

$$\mathbb{E}_{Cr^*}\mathcal{I}(Cr^*, \mathcal{F}, \cdot) = \sum_{i=1}^{N} w_i \mathbb{E}_{Cr_i}\mathcal{I}(Cr_i, \mathcal{F}, \cdot), \tag{22}$$

where $\{w_1, \ldots, w_N\}$ is a set of positive weights that sum to 1. Under these conditions, the credence function that minimizes the group's expected inaccuracy is necessarily a linear mean of individual credences. Thus, Moss argues, there are accuracy-based reasons to

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

solve the group credence problem by taking a linear mean of individual credences. In response, I note that Moss assumes in her argument that the expected accuracy of a group credence, according to that group credence, must be a linear mean of the expected accuracy of each individual credence, according to those credences. This assumption closes off the possibility of group credences that have greater expected accuracy, according to some true probability distribution, in cases where all individuals have highly inaccurate credences. Thus, Moss's crucial assumption is not motivated by the norm that the group ought to minimize expected inaccuracy in all cases, where expected inaccuracy is measured by an objective standard rather than by the subjective standards of the individuals in the group.

In summary, I am committed in this article to the following view: the norm that individuals in a group ought to plan to endorse group credences that they subjectively expect to be accurate is only justified on veritistic grounds to the extent that their future subjective beliefs will be good approximations of what is in fact the case in the actual world. In the scenario that I considered in the previous section, individuals' credences were derived from models that, by stipulation, were not good approximations of what was in fact the case in the actual world. Thus, in such a scenario, it does not follow that individuals ought to endorse group credences that they expect to be accurate, at least where the *ought* in "ought to endorse" is interpreted in an externalist, objective sense rather than an internalist or subjective sense. That is, if we take the primary norm of epistemology to be that agents ought to have beliefs that are close to the truth, rather than having beliefs that *they believe* will be close to the truth, then the putatively accuracy-based arguments of Pettigrew (2019) and Moss (2011) in favor of linear averaging do not go through. By contrast, the same externalist understanding of accuracy-based norms for epistemology does motivate a stacking-based solution to the group credence problem, as demonstrated by the results in the previous section.

6 Counterarguments

A key objection to my argument so far proceeds as follows: Applying stacking methods to solve the group credence problem requires that all agents in a group have access to a shared data set and have a fairly sophisticated mechanism for analyzing the data and returning a hypothesis that describes the data-generating process. Further, each agent must be able to perform this level of data analysis not just when given the full data set but also when given each data set that can be generated by removing one of the data points from the full set. These facts about stacking present two ways of arguing that stacking cannot be used to solve the group credence problem. First, one could argue that the presupposition that agents could derive a stacking-based solution to the group credence problem endows those agents with an unrealistic level of epistemic ability. Second, one could argue that in many cases, groups of agents will lack access to a shared data set that they can analyze to produce models and derive credences in future events. Rather, it could be argued, in many cases, agents' credences are simply opinions that are not derived from data analysis.

Regarding the first counterargument, I need only point out that any formal approach to the group credence problem that assumes that each individual in the group assigns a specific numerical credence to each element of a partition, and indeed, each element of an algebra over a set of possible worlds, already represents

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

agents in ways that idealize away from actual epistemic life. Real-world agents do not come with well-defined credences in an exhaustive set of possible events, and attempts to elicit such credences via iterated gambles are unlikely to be feasible in practice and may nevertheless be undermined by inconsistent betting behavior on the part of real-world agents. Further, groups of agents rarely, if ever, possess a mutually agreed-upon set of possible worlds to serve as the sample space over which an algebra is defined. These facts render the very framing of the group credence problem an idealization. None of this implies that formal epistemology cannot be useful in providing normative guidance on how to solve the group credence problem. Rather, just as scientific models idealize away from their target systems while still providing important insights about the nature of those systems, formal epistemology can provide insight into epistemic practice while nevertheless presenting an idealized picture of actual epistemic practice. Thus, although stacking introduces further idealized elements into the group credence problem, the introduction of said elements does not necessarily imply that stacking is not applicable to the group credence problem.

As for the second counterargument, I hold that when the credences of individuals in a group are mere opinions, not based on any analysis of underlying data, then we have no reason to suspect that their group credences should be highly accurate.[4] To illustrate why this is the case, consider the earlier example in which Alphonse and Belinda disagree about whether the train to Brussels leaves before or after noon. If neither has any data regarding the schedule of trains from Brussels to Amsterdam (e.g., neither has looked at a schedule, neither has taken the train before, neither has any experience with train journeys between European capitals, etc.), then there is no reason why we should expect either of their credences to have any probative value with respect to the time at which the train is likely to depart. In these kinds of cases, the epistemologist who takes accuracy to be the sole source of value for group credences has no reason to suspect that *any* mathematical operation combining Alphonse and Belinda's two credences into a single credence will get them significantly closer to the truth. By contrast, if Alphonse and Belinda *do* have access to some of the data sources described earlier, then their process of arriving at a group credence can be represented, with some idealization, as a stacking-based aggregation of models from which a group credence can be derived.

This discussion reveals what I believe is an important epistemic upshot of stacking-based solutions to the group credence function. Epistemologists who argue that the sole source of epistemic value for a credence function is the accuracy of that credence function are what Berker (2013) calls "epistemic consequentialists." Just as consequentialists in ethics believe that the moral valence of an action is determined solely by its consequences and not an agent's reasons for performing that action, those who hold up accuracy as the sole epistemic virtue of a credence function believe that the epistemic valence of a credence function is measured solely by the extent to which that credence function allows an agent to believe the truth. Importantly, the agent's *reasons* for adopting partial beliefs consistent with that credence function are

[4] A result from Rougier (2016) does show that, in general, an average of credence functions in a given set has greater expected accuracy than a randomly selected credence function from that same set. However, this does not establish that a mean credence function cannot be significantly less accurate than some other credence function not in the set being averaged.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

irrelevant. Berker, for his part, rejects epistemic consequentialism, insisting that the epistemic value of a belief is determined at least in part by an agent's reasons for holding that belief.

What stacking-based approaches to the group credence problem show is that even if one believes that accuracy is the sole source of value for a credence function, one should still care whether agents in a group have reasons for adopting partial beliefs that are consistent with a particular credence function. In the stacking cases, each agent has a particular analysis of shared data on which they base their credences. Thus, they have a reason for holding the partial belief that they do. Even if an agent does not have a *good* reason for holding their partial beliefs (e.g., they are a poor data analyst), making these reasons explicit to the group can facilitate stacking, which in turn allows for a more accurate solution to the group credence problem than would have been possible had each individual provided only credences and no reasons justifying their holding those credences. Thus, having a reason for holding a given credence is better than holding the same credence for no reason at all, insofar as one is a member of a group that wishes to have an accurate group credence function. This offers a possible point of conciliation between epistemic consequentialists and their rivals. Both camps can agree that when each individual agent in a group has a reason for holding partial beliefs that cohere with a given credence function, the existence of said reasons can improve the value of the group's epistemic state. However, they disagree over whether this improvement is directly due to the existence of said reasons or due to the increased accuracy of the group credences that these reasons enable.

An additional counterargument against what I have presented here could proceed as follows: In the example that I have given previously, Alphonse and Belinda's learning algorithms produce such inaccurate models that they must either possess evidence not represented in their mutual data set or else they must be irrational in some sense. Let us stipulate that neither Alphonse nor Belinda has any special knowledge that affects the output of their learning algorithm. This leaves open the possibility that the severe inaccuracy of Alphonse and Belinda's learning algorithms is due to putative irrationality on their part. If this is the case, the counterargument might continue, then the failure of any linear mean to produce an accurate group credence is explained not by any flaw in the method of taking a linear mean of individual group credences but instead by Alphonse and Belinda's irrationality.

In response, I note first that I am wary of equating severe inaccuracy with irrationality. Because the inaccuracy of a prediction comes in degrees, the claim that a prediction can be so inaccurate that it renders the predicting agent irrational invites a version of the Sorites paradox. If a severely inaccurate prediction ρ_1 renders the predictor irrational, then so, too, it stands to reason, does a prediction ρ_2 that is only slightly more accurate than ρ_1, and so on until all predictors are declared irrational. I note, second, that although my previous example involved severely inaccurate predictors in order to provide a clear demonstration of the virtues of stacking, stacking can still outperform credal averaging when predictors are less severely inaccurate. To illustrate, consider the same example as previously noted but with Alphonse and Belinda's models changed to the following, whereas the true model $M_T : h = 5l + \epsilon$ remains unchanged:

$$M_A : h = 6l + \epsilon. \tag{23}$$

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

$$M_B : h = 6.1l + \epsilon. \tag{24}$$

Thus, Alphonse and Belinda both overestimate the number of hares that ought to be present in the region, given the number of lynxes, but their inaccuracy is less egregious than in the earlier example. Suppose further that Alphonse and Belinda's leave-α-out vectors are as follows: $\vec{z}_A = [6, 12, 17, 23]$, $\vec{z}_B = [7, 13, 18, 24]$. Applying stacking in this case yields the following aggregate model:

$$M_S : h = .81(6l) + .09(6.1l) + \epsilon = 5.45l + \epsilon. \tag{25}$$

If this stacked model is used to generate group credences, then the group's credence function over the number of hares observed on a given day will be centered around a mean that is closer to the mean predicted by the true model than the mean around which either Alphonse's or Belinda's credence function is centered. Indeed, both Alphonse and Belinda's credence functions will be centered around means that are *greater* than the mean around which the true probability distribution will be centered. Thus, stacking yields a more accurate group credence than would be possible under linear averaging, even when the inaccuracy of the individuals in the group is less severe than in the earlier case.

Another objection might be to take issue with my stipulation of a true data-generating process in the lynx–hare scenario, according to which the expected accuracy of each group credence is calculated. According to this objection, although it is true that the data are made more likely by the data-generating process, such that $h = 5l + \varepsilon$ is statistically better motivated as the true model compared with Alphonse's model $h = 2l + \varepsilon$ or Belinda's model $h = 3l + \varepsilon$, there is no logical barrier to either Alphonse's or Belinda's model, or some linear mean of their models, representing the true data-generating process. In the unlikely but not impossible event that this is the case, stacking will produce a *less* accurate model than taking a linear mean. In response, I note that in light of the result summarized in equation (14), in the long run of data collection, we expect the stacked model to be highly accurate and thus to generate highly accurate credences, regardless of the specific form of the true data-generating process. Of course, it can occur that small data sets, such as the one that I have used in my example for the sake of exposition, are highly improbable within the context of a given data-generating model, such that algorithms trained on small samples give poor results. However, the epistemic norm that a group ought to behave in a way that maximizes the long-run expected accuracy is part of the underlying motivation of both my approach and the approach of defenders of the linear mean, such as Moss or Pettigrew. From this, I conclude that the threat of inaccurate predictions due to unlikely outcomes in finite sampling is no more a problem for my view than it is for their views—or indeed, for any expected-accuracy–motivated argument in favor of groups or individuals adopting a particular set of beliefs in response to observed data.

A final objection against what I have presented here proceeds as follows: In any specification of the group credence problem such that stacking-based solutions are appropriate, there is a shared data set to which all individuals in the group have access. In cases where all individuals have highly inaccurate learning algorithms, why not solve the group credence problem by simply using a better learning algorithm than that possessed by any individuals in the group to build a new model of

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

the data-generating process and then derive the group credence from that model? Such a credence could have greater expected accuracy than a credence derived from the stacking-based aggregation of the two inaccurate models. Thus, it could be argued, simply training a different learning algorithm on the commonly available data can, in some cases, be a better method for solving the group credence problem, from a veritistic perspective, than stacking the inaccurate models. In response to such an objection, I need only note that once we introduce a better learning algorithm into the group credence problem, we are effectively introducing a new, more accurate individual into the group. By virtue of the results stated previously, we can still expect the model produced by stacking the accurate and inaccurate models to be as or more accurate, in the long run, than the single inaccurate model. Thus, stacking is to be preferred to any single learning algorithm in producing a solution to the group credence problem.

7 Conclusion

I have presented the group credence problem, now the subject of considerable attention in formal social epistemology. I have shown how, when all members of a group are highly inaccurate, any solution to the group credence problem that relies on taking a mean of each individual in the group's credence will result in an inaccurate group credence. I then show how, if we model each agent's credence function as derived from a learning algorithm trained on a common data set, we can use stacking techniques to improve the accuracy of the group's credence function. In so doing, I reject Pettigrew's (2019) unanimity constraint on any solution to the group credence problem, arguing that there is no accuracy-based reason for accepting the constraint. I defend this approach against potential counterarguments, showing that even if reasons for belief are not in themselves sources of value for a set of partial beliefs, epistemologists concerned with the accuracy of group credences should seek to solicit the reasons for the partial beliefs from the individuals in a group. Even where all individuals in a group have bad reasons for holding the credences that they do, such an elicitation of reasons can facilitate stacking, an aggregation method that yields more accurate group credences than averaging techniques in some cases.

It is worth clarifying that I do not take stacking to be a panacea for the group credence problem. In particular, the law-of-large-numbers justification of stacking does not work in cases where exchangeability is violated (i.e., in cases where data are more likely to appear in a specific order than in some permutation of that order). It also is not applicable, at least not in the form presented here, when individuals possess private data in addition to the public data available to the entire group or when the group does not have access to enough data. Finally, it may be that the predicted value of interest is an element of a different domain than the data values or that the data-generating process cannot be assumed to be consistent from past to future. Here, too, the stacking methods presented may not be applicable. Nevertheless, I take myself to have shown that in at least some important cases, stacking is able to generate more accurate group credences than mean-based aggregation methods. Moreover, in the cases in which stacking will not produce an accurate solution, there is no reason to think that mean-based aggregation methods *will* produce an accurate prediction.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

That said, the method of stacking presented here is only one version of the stacking methodology; for a sampling of the many variants of stacking on offer, see Breiman (1996b), Van der Laan, Polley, and Hubbard (2007), Sill, Takacs, Mackey, and Lin (2009), and Clyde and Iversen (2013). For an implementation of "hierarchical stacking," in which stacking weights can vary across a population of learning algorithms and in which stacking weights can vary as a function of continuous predictors, see Yao, Pirs, Vehtari, and Gelman (2021). Moreover, there are other popular model-aggregation procedures from statistics and machine learning that may prove effective in various contexts. These include boosting (for a textbook treatment, see Schapire and Freund 2013), bagging (Breiman 1996a), a mixture of experts (see Yuksel, Wilson, and Gader [2012] and Masoudnia and Ebrahimpour [2014] for surveys), and prequential analysis (Dawid 1984). Examining how each of these techniques can be deployed in various contexts to achieve a satisfying solution to the group credence problem is an intriguing avenue for future research at the intersection of machine learning, statistics, and social epistemology.

References

Berker, Selim. 2013. "The Rejection of Epistemic Consequentialism." *Philosophical Issues* 23 (1):363–87.

Bernardo, José M., and Adrian F. M. Smith. 1994. *Bayesian Theory*. New York: John Wiley & Sons.

Bradley, Richard. 2007. "Reaching a Consensus." *Social Choice and Welfare* 29 (4):609–32.

Breiman, Leo. 1996a. "Bagging Predictors." *Machine Learning* 24 (2):123–40.

Breiman, Leo. 1996b. "Stacked Regressions." *Machine Learning* 24 (1):49–64.

Clyde, Merlise, and Edwin S. Iversen. 2013. "Bayesian Model Averaging in the M-Open Framework." In *Bayesian Theory and Applications*, edited by P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens, 483–98. Oxford: Oxford University Press.

Dawid, A. Philip. 1982. "The Well-Calibrated Bayesian." *Journal of the American Statistical Association* 77 (379):605–10.

Dawid, A. Philip. 1984. "Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach." *Journal of the Royal Statistical Society: Series A (General)* 147 (2): 278–90.

Dietrich, Franz. 2019. "A Theory of Bayesian Groups." *Noûs* 53 (3):708–36.

Dietrich, Franz, and Christian List. 2017. "Probabilistic Opinion Pooling Generalized. Part One: General Agendas." *Social Choice and Welfare* 48 (4):747–86.

Goldman, Alvin. 2001. "The Unity of the Epistemic Virtues." In *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*, edited by A. Fairweather and L. Zagzebski, 30–48. Oxford: Oxford University Press.

Jehle, David, and Branden Fitelson. 2009. "What Is the 'Equal Weight View'?" *Episteme* 6 (3):280–93.

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

Kadane, Joseph B., and Sarah Lichtenstein. 1982. *A Subjectivist View of Calibration*. Technical Report. Springfield, OR: Decision Research.

Kuan, Ko-Hung. 2021. "Beyond Linear Conciliation." *Synthese* 198 (12):11483–504.

Le, Tri, and Bertrand Clarke. 2017. "A Bayes Interpretation of Stacking for $\mathcal{M}$-Complete and $\mathcal{M}$-Open Settings." *Bayesian Analysis* 12 (3):807–29.

Lehrer, Keith, and Carl Wagner. 1983. "Probability Amalgamation and the Independence Issue: A Reply to Laddaga." *Synthese* 55 (3):339–46.

Masoudnia, Saeed, and Reza Ebrahimpour. 2014. "Mixture of Experts: A Literature Survey." *Artificial Intelligence Review* 42 (2):275–93.

Moss, Sarah. 2011. "Scoring Rules and Epistemic Compromise." *Mind* 120 (480):1053–69.

Pettigrew, Richard. 2019. "On the Accuracy of Group Credences." In *Oxford Studies in Epistemology*, vol. 6, edited by T. S. Gendler and J. Hawthorne, 137–60. Oxford: Oxford University Press.

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press

Rougier, Jonathan. 2016. "Ensemble Averaging and Mean Squared Error." *Journal of Climate* 29 (24): 8865–70.
Russell, Jeffrey Sanford, John Hawthorne, and Lara Buchak. 2015. "Groupthink." *Philosophical Studies* 172 (5):1287–1309.
Schapire, Robert E., and Yoav Freund. 2013. *Boosting: Foundations and Algorithms*. Bingley, UK: Emerald Group.
Sill, Joseph, Gábor Takács, Lester Mackey, and David Lin. 2009. "Feature-Weighted Linear Stacking." arXiv preprint, arXiv:0911.0460.
Staffel, Julia. 2015. "Disagreement and Epistemic Utility-Based Compromise." *Journal of Philosophical Logic* 44 (3):273–86.
Steele, Katie. 2012. "Testimony as Evidence: More Problems for Linear Pooling." *Journal of Philosophical Logic* 41 (6):983–99.
Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1):article 25.
Wagner, Carl. 1985. "On the Formal Properties of Weighted Averaging as a Method of Aggregation." *Synthese* 62 (1):97–108.
Wolpert, David H. 1992. "Stacked Generalization." *Neural Networks* 5 (2):241–59.
Yao, Yuling, Gregor Pirš, Aki Vehtari, and Andrew Gelman. 2021. "Bayesian Hierarchical Stacking." arXiv preprint, arXiv:2101.08954.
Yuksel, Seniha Esen, Joseph N. Wilson, and Paul D. Gader. 2012. "Twenty Years of Mixture of Experts." *IEEE Transactions on Neural Networks and Learning Systems* 23 (8):1177–93.

Cite this article: Kinney, David. 2022. "Why Average When You Can Stack? Better Methods for Generating Accurate Group Credences." *Philosophy of Science* 89 (4):845–863. https://doi.org/10.1017/psa.2022.3

https://doi.org/10.1017/psa.2022.3 Published online by Cambridge University Press