MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems

Tom Wilkening
Department of Economics, The University of Melbourne, Parkville VIC 3010 tom.wilkening@unimelb.edu.au

Marcellin Martinie, Piers D. L. Howe
Melbourne School of Psychological Sciences, The University of Melbourne, Parkville VIC 3010 m.martinie@unimelb.edu.au
pdhowe@unimelb.edu.au

Modern forecasting algorithms use the Wisdom of Crowds to produce forecasts better than those of the best identifiable expert. However, these algorithms may be inaccurate when crowds are systematically biased or when expertise varies substantially across forecasters. Recent work by Prelec et al. (2017) has shown that meta-predictions—a forecast of the average forecast of others—can be used to correct for biases even when no external information such as forecasters' past performance is available. We explore whether meta-predictions can also be used to improve forecasts by identifying and leveraging the expertise of forecasters. We develop a confidence-based version of the Surprisingly Popular algorithm of Prelec et al. (2017). Like the original algorithm, our new algorithm is robust to bias. However, unlike the original algorithm, our version is predicted to always weight forecasters with more informative private signals more than forecasters with less informative ones. In a series of experiments, we find that the modified algorithm does a better job in weighting informed forecasters than the original algorithm and show that individuals who are correct more often on similar decision problems contribute more to the final decision than other forecasters. Empirically, the modified algorithm outperforms the original algorithm for a set of 500 decision problems.

Key words: expertise, meta-knowledge, wisdom of crowds, forecasting, aggregation
History: This paper was first submitted on 21 February 2020.

1. Introduction

The Wisdom of Crowds has revolutionized the way in which we make predictions. It is the phenomenon where crowds make consistently better predictions, judgments, or estimates than even the most-expert individuals (Galton 1907, Surowiecki 2005). The superiority of aggregate predictions over individual predictions has been demonstrated across a variety of domains, but has gained particular attention in economic, political, and market forecasting where there are often high stakes involved (Budescu and Chen 2015, Dreber et al. 2015, Mellers et al. 2015, Müller-Trede et al. 2017, Tetlock 2017, Gillen et al. 2018).

The simplest approach to aggregating predictions is to use majority voting. As shown by the Condorcet Jury Theorem (Condorcet 1785), the probability that majority voting produces the correct decision for a binary decision increases towards 100% as the group sizes increases, under the assumption that each individual is more likely to be correct than incorrect. Despite its appealing properties, majority voting may often be inaccurate when crowds contain a large proportion of uninformed forecasters or when the population of forecasters are systematically biased (Simmons et al. 2011, Budescu and Chen 2015).

To deal with the issue of uninformed forecasters, researchers have developed aggregation techniques that use training data to identify and weight forecasters based on their expertise. For example, Cooke (1991) developed a model that identifies and excludes non-experts from the crowd based on their performance on seed questions with outcomes that are known to the decision-maker. Similarly, Budescu and Chen (2015) showed that significant improvements in accuracy over the unweighted mean could be obtained by weighting experts by their performance relative to the crowd and excluding forecasters who did not improve the aggregate prediction.

Although expert-selection methods often generate better predictions than majority voting, researchers are not always able to identify individuals with the relevant expertise in advance. For example, forecasters' performance on prior problems with known outcomes might not effectively predict performance on problems of actual interest, and collecting the responses to a panel of relevant problems may be impractical (Genre et al. 2013, Clemen 1989). We refer to forecasting problems where it is either not possible or not helpful to use the individual's responses to prior problems as "single-question" forecasting problems, as the task is then to make the best forecast possible based on data relating only to a single question. We concentrate on the single-question problem for the rest of the paper.

The standard approach to the single-question prediction problem has been to use reported confidence to weight forecasters or to simply select the answer with the highest confidence (Koriat 2012, Prelec et al. 2017). These confidence-based approaches treat confidence as a predictor of expertise, weighting more-confident judgments more than less-confident judgments in the aggregation process. However, forecasters who hold the majority opinion tend to be overconfident while individuals who hold the minority opinion tend to be under-confident (Hertwig 2012, Koriat 2008, 2012). Thus, confidence may be negatively correlated with accuracy in "wicked" problems where most forecasters are incorrect. Indeed, there are many examples in the literature in which incorrect forecasters are more confident in their forecasts than correct forecasters (Koriat 2008, 2012, Fischhoff and MacGregor 1982, Lee and Lee 2017).

In this paper we explore whether we can improve upon existing confidence-based approaches by combining forecasts with meta-predicitons about the forecasts of others. In a remarkable paper,

Prelec et al. (PSM, 2017) proposed a novel algorithm that uses meta-predictions to correct for crowd biases. Their Surprisingly Popular (SP) algorithm generates predictions by using forecasters' votes about whether a particular event will be true or false and forecasters' meta-predictions—a prediction of the proportion of other forecasters that will vote true. The SP algorithm predicts the outcome that is more popular than the crowd expects (i.e., the surprisingly popular outcome) to be the correct answer. In other words, the SP algorithm predicts true when the total number of true votes exceeds the average of the meta-predictions, and false otherwise.

PSM showed that the SP algorithm has the important theoretical property that it will always predict the correct answer when aggregating reports from a large homogeneous population of Bayesian forecasters, even when a substantial fraction of these forecasters are biased. In the first section of this paper, we show that an alternative Surprisingly Confident (SC) algorithm, which generates predictions by using forecasters' confidences and meta-predictions about the confidences of others, also shares this property. We then explore the theoretical properties of the SP and SC algorithms as they relate to expertise.

In our theoretical framework, we consider an environment in which individuals are asked binary true or false problems and share a common prior about the likelihood that the answer is true. Individuals receive signals from an information service and form a posterior belief about whether the answer is true using Bayes rule. The posterior belief held by an individual influences both their vote and their meta-prediction of the votes of others. We say that an individual has *a more informative private signal* than another individual if (i) the two individuals have posteriors that are both above the common prior or below the common prior and (ii) the absolute distance between the first individual's posterior and the common prior is larger than the second. An algorithm leverages informed forecasters if individuals with more informative signals contribute more to the algorithm's final prediction than those who are uninformed.

Our first result is that the SP algorithm actually leverages *uninformed* forecasters in problems where the crowd is initially unbiased. That is, the contribution that an individual makes to the final prediction of the algorithm is decreasing in the quality of the individual's information, such that individual forecasters who receive the most information have lower contributions to the aggregated forecast than individual forecasters who receive less information. To prove this result, we provide a novel alternative formulation of the SP algorithm, which expresses the algorithm in terms of a weighted average of the forecasters' votes. In this formulation, the weight assigned to an individual's vote is proportional to the absolute difference between a forecasters vote and their meta-prediction about the vote of others. We show that in unbiased problems, the weights are largest for fully uninformed individuals and strictly decrease as an individual becomes better informed.[1]

[1] This result does not mean that the SP algorithm will always perform poorly in unbiased settings. We show in our examples that the SP algorithm reveals important information that is common knowledge to all forecasters regarding

Our modified SC algorithm improves on the way the algorithm weights forecasters with better-informed signals. Specifically, the SC algorithm always leverages forecasters with more informative private signals regardless of whether the decision problem is biased or unbiased. Thus, a forecaster who has a more informative signal will always make a larger contribution to the final outcome of the algorithm than one who has a less informative signal. We show that the differences in the weight functions between the SP and SC algorithm may lead the SC algorithm to be more accurate than the SP algorithm when the sample of forecasters is finite, particularly in cases where forecasters vary in expertise.

Although our first result suggests that the SP algorithm may over-weight uninformed individuals and under-weight informed ones, it ignores a key advantage of crowd forecasts. In problems with an unbiased prior, the votes of forecasters who receive no information will be random while the votes of those who know the correct state will be perfectly correlated. This will cause the votes of the uninformed forecasters to partially cancel out as crowd size increases and may offset the weighting of individuals.

To understand the aggregate properties of the SP and SC algorithms, we consider a more general environment in which individuals share the same prior belief but have access to one of two information services that are ordered in terms of informativeness. We refer to individuals who draw signals from the more informative information service as experts and individuals who draw signals from the less informative service as novices. Although experts and novices are assumed to have the same priors, experts are expected, on average, to receive more informative private signals and therefore predict the correct answer more often than novices. An algorithm leverages this expertise if the expected contribution of an expert is greater than that of a novice in both true and false questions.

As a second result, we show that the SC algorithm will leverage expertise in any environment where private signals are independent after conditioning on the state.[2] By contrast, the SP algorithm requires additional assumptions to ensure that the algorithm leverages expertise. In Appendix B, we derive a set of sufficient conditions on the structure of the information services that guarantee that the SP algorithm leverages expertise. Our conditions suggest that in unbiased problems, experts will be leveraged by the SP algorithm in environments where (i) there is a mix of both experts and novices in the population and (ii) novices are reasonably uninformed.

the structure of signals. In large samples, this information is enough to correctly predict the right answer in both biased and unbiased problems when all forecasters know the true distribution of potential signals.

[2] The SC algorithm is able to leverage expertise in cases where individuals share a biased common prior. Such priors may come from a commonly observable public signal. Thus the algorithm will also leverage expertise in environments where individuals receive both a commonly observed public signal and conditionally independent private signals. See Palley and Soll (2018) for an alternative probabilistic forecasting algorithm that is designed to account for more complex signal correlation structures.

Finally, we consider the properties of the SC algorithm in more realistic settings where reported confidences do not coincide with each forecaster's posterior and where forecasts are systematically miscalibrated. We show that even when forecasters are not Bayesian, the SC algorithm will predict the true answer in large samples if (i) reported confidences are weakly increasing in the underlying true posteriors and (ii) forecasters take systematic overconfidence into account when reporting their meta-prediction. This result suggests that the algorithm is likely to perform well in settings where forecasters who believe the consensus position is correct are overconfident and forecasters who believe the consensus position is incorrect are under-confident. In such environments, other confidence-based aggregation approaches tend to fail.

Our theoretical results predicts particular patterns in the weights generated in the SP algorithm that vary with initial crowd bias. To analyse whether these patterns exist empirically, we estimate the relationship between weights and signals in two datasets: a replication of the US States dataset of PSM in which the prior is predicted to be strongly biased, and a new quiz dataset where we can vary the distribution of experts and novices by varying task difficulty. Using the probabilistic forecasts of an individual as a proxy for their posterior belief, we show that the weights in the datasets from both our experiments follow the patterns predicted by the theory for both the SP algorithm and the SC algorithm.

Our theoretical model also predicts that in unbiased problems, the SP algorithm is likely to perform well when there is variation in experts and non-experts in the environment. To test for this feature, we systematically vary the difficulty of problems in our new quiz dataset to create variation in problem difficulty and the likely mix of novices and experts. Consistent with our theoretical predictions, the SC algorithm leveraged expertise more effectively than the SP algorithm for all difficulty levels.

Finally, we compare the performance of the SC and SP algorithms across our two experiments. We find that the SC algorithm outperforms the SP algorithm in our quiz datasets and that this outperformance is driven by difficult problems where the SC algorithm performs well. Surprisingly, the SC algorithm performs poorly in easy quiz problems. We discuss how this may be due to the treatment of commonly observed signals in the SC algorithm and briefly discuss how the SC and SP algorithms might be combined to improve forecasts over each algorithm on its own.

Our paper contributes to the literature by providing a single-question algorithm that has promising empirical and theoretical properties in terms of expertise. The SC algorithm is robust to bias and corrects for overconfidence in situations where other confidence-based aggregation approaches

fail. Further, under reasonable assumptions, the SC algorithm has the intuitive feature that uninformed individuals will be given zero weight and maximally informed individuals will be given the highest weight.[3]

The rest of the paper is structured as follows. We present our main theoretical results in Section 2 and test these results empirically in Section 3. We collect all proofs for the lemmas and propositions in Appendix C.

2. Theory

We consider a Bayesian model in which a crowd of N forecasters is assembled to predict the outcome of a single event. The outcome of the event, $o \in \{T, F\}$, is binary and can be true or false. Forecasters share a common prior $p(T)$ that the event is true.

Each forecaster receives a private signal S, that is a random variable taking on real value realisations in the set $\{s_1, \ldots, s_m\} \cup \{s_\varnothing\}$ where $0 \leq s_1 < s_2 < \cdots < s_m \leq 1$ and $s_1 < s_\varnothing < s_m$. As our outcome space is binary, it is without loss of generality that we normalise the signals so that their value is equal to the posterior belief that an event is true. That is, $s_j := p(T|s_j)$. We let $s_\varnothing$ represent the case where an individual receives an uninformative signal so that $s_\varnothing := p(T)$.

To minimise ambiguity, we will use s_j to denote the jth lowest posterior in the set $\{s_1, \ldots, s_m\} \cup \{s_\varnothing\}$. Thus, it is always the case that $s_1 < s_2$. We will use σ_k to denote the signal drawn by a particular forecaster k. As each signal is drawn randomly, there is no inherent order between σ_1 and σ_2.

We use a left stochastic matrix called an *information service* to model the distribution of signals across forecasters in each state.[4] Initially, we will assume that all participants receive signals from the same information service denoted as Q.[5] We also assume that the properties of Q are common knowledge to all forecasters.

An information service is composed of a likelihood matrix $[Q_{oj}]_{2\times(m+1)}$. Each element of the first row of Q represents the probability that the signal is s_j given the outcome is $o = T$. Likewise, each

[3] The weights used in the SC algorithm can also be used in probabilistic forecasting problems. See Martinie et al. (2020) for a discussion of how the weights of the SC algorithm can be adapted to the probabilistic forecasting domain and for a comparison of the algorithm to other probabilistic forecasting algorithms proposed by Palley and Soll (2018) and Satopää et al. (2016). McCoy and Prelec (2017) develops an alternative Bayesian hierarchical model that can be used in forecasting problems with multiple-choice answers. A distinguishing feature of the model is that it works in both the single-question and multiple-question domains. Palley and Satopää (2020) propose a selection criterion that excludes forecasters with inaccurate meta-predictions from the aggregated crowd forecast. Their selection approach adjusts the number of experts selected based on the level of noise observed in the forecasters' predictions of others and is particularly well suited to situations where meta-predictions have an unknown level of noise.

[4] See Blackwell (1953), Blackwell and Girshick (1979), Marschak and Miyasawa (1968), Marschak and Radner (1972) for general treatments of information services.

[5] In Subsection 2.3, we will relax this assumption and introduce experts who will receive signals from a more informative information service.

element of the second row of Q represents the probability that the signal is s_j given the outcome is $o = F$. For ease, we will denote the first row elements with T and the second row elements with F. Thus $Q_{Tj} := Q_{1j} = p(s_j|T)$ while $Q_{Fj} := Q_{2j} = p(s_j|F)$.

We note two important features of an information service. First, an information service acts as a transition matrix from a state of nature to a signal and thus $\sum_j Q_{oj} = 1$ for each row $o \in \{T, F\}$. Second, upon receiving a message from an information service, agents revise their priors using Bayes rule. For any signal that occurs with positive probability (i.e., where $Q_{Tj} + Q_{Fj} > 0$), the posterior belief that the event is true is given by

$$p(T|s_j) = \frac{p(T)Q_{Tj}}{p(T)Q_{Tj} + p(F)Q_{Fj}}.$$

By construction, this is equal to s_j for all signals that occur with positive probability.

It will be useful to classify decision problems based on the properties of Q. The following definitions help to identify three types of decision problems, which will respond differently across aggregation problems. We first classify decision problems based on whether the common prior is biased or unbiased:

Definition 1. A decision problem has an **unbiased prior** if $s_\varnothing = 0.5$ and a **biased prior** if $s_\varnothing \neq 0.5$.

We further divide the class of unbiased problems into asymmetric and symmetric decision problems. We will call an information service *symmetric* if the likelihood of posterior s_i in state T is equal to the likelihood of posterior $(1 - s_i)$ in state F. Symmetry places restrictions both on the set of outcomes and on the relationship between likelihoods.

Definition 2. An information service is **symmetric** if (i) $s_\varnothing = \frac{1}{2}$, (ii) the cardinality of the set $\{s_1, \ldots, s_m\}$ is even, and (iii) $Q_{Ti} = Q_{F(m-i+2)}$ for all $i \in \{1, \ldots, m+1\}$.

Following Prelec et al. (2017), we will focus attention to information services that have the following property:

Definition 3. An information service Q is **responsive** if there is a positive probability that a forecaster votes for the correct answer both when the state is true and when the state is false:

$$\sum_{\{i|s_i \leq .5\}} Q_{Fi} > 0 \quad \textit{and} \quad \sum_{\{i|s_i \geq .5\}} Q_{Ti} > 0.$$

Responsive information services require that the bias is not so strong that all forecasters will go against their own private information and vote with the publicly observable signal in large samples.

The assumption will imply that the expected vote in the true state is larger than the expected vote in the false state.

Finally, we will use the following partial ordering of signals to evaluate how the algorithm treats individuals with different amounts of information.

DEFINITION 4. Forecaster i has a **more informative private signal** than forecaster j if either (i) $\sigma_i < \sigma_j < s_\varnothing$ or (ii) $\sigma_i > \sigma_j > s_\varnothing$.

Intuitively, the informativeness of a forecasters private signal is related to the distance between his posterior and the common prior. We have restricted attention to cases where σ_i and σ_j are either both greater than $s_\varnothing$ or both less than $s_\varnothing$ so that distance is directly related to the relative changes in the likelihood ratios of the two forecasters.[6]

We note that the ordering of private signals is related to the extremity of the posterior, but is not equivalent to extremity in decision problems where there is a biased priors. For example, in a problem where the common prior is $s_\varnothing = .75$, a forecaster who has a signal of $\sigma_i = 0.5$ will have received a more informative signal than a forecaster with a signal of $\sigma_j = 0.6$.

2.1. Single-question forecasting algorithms

We consider single-question forecasting algorithms that use information from predictions and meta-predictions about the current event only. Let $V_i(T|\sigma_i) \in \{0,1\}$ be the forecaster's prediction, or vote, that the event is true given signal σ_i, and let $P_i(T|\sigma_i) \in [0,1]$ be the forecaster's probabilistic forecast that the event is true. Further let $M_i^V(Q|\sigma_i) \in [0,1]$ be a forecaster's **vote meta-prediction**: a forecaster's meta-prediction of the share of other forecaster's that will vote true. Let $M_i^P(Q|\sigma_i) \in [0,1]$ be a forecaster's **probability meta-prediction**: the forecaster's meta-prediction of the average probability forecast of all other forecasters. To simplify notation, we let $V_i := V_i(T|\sigma_i)$, $P_i := P_i(T|\sigma_i)$, $M_i^V := M_i^V(Q|\sigma_i)$, and $M_i^P := M_i^P(Q|\sigma_i)$.

We let $X_i := (V_i, P_i, M_i^V, M_i^P)$ be forecasters i's full report and let $X = (X_1, X_2, \ldots, X_N)$ be the full reports of all forecasters. Each algorithm we consider is a mapping $T : X \to \{0,1\}$, which aggregates the data from a single event into a categorical forecast of whether the event is true or false. We assume the forecasters are truthful in all the algorithms and that they randomise their votes uniformly if they have the *uninformed* posterior of 0.5. This implies that $V_i = 0$ if $P_i < 0.5$, $V_i = 1$ if $P_i > 0.5$, and V_i is equally likely to be zero or one when $P_i = 0.5$.

We explore the theoretical properties of two alternative meta-prediction algorithms in this paper: the Surprisingly Popular (SP) algorithm of Prelec et al. (2017) and a variant that we refer to as the Surprisingly Confident (SC) algorithm. In the SP algorithm, the proportion of the crowd

[6] For example, if $s_i > s_j > s_\varnothing$, then $\frac{Q_{Ti}}{Q_{Fi}} > \frac{Q_{Tj}}{Q_{Fj}} > \frac{p(T)}{p(F)}$. Thus $|s_i - s_\varnothing| > |s_j - s_\varnothing|$ implies $|\frac{Q_{Ti}}{Q_{Fi}} - \frac{p(T)}{p(F)}| > |\frac{Q_{Tj}}{Q_{Fj}} - \frac{p(T)}{p(F)}|$.

voting true is compared to the mean vote meta-prediction. If the proportion of true votes exceeds the average of the vote meta-prediction, the event is predicted to be true. Otherwise, the event is predicted to be false. Formally,

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N}(V_i - M_i^V) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Although the standard formulation of the SP algorithm is easy to compute, it is relatively difficult to understand how the algorithm treats forecasters with different signals. The following lemma provides an alternative "weighted average" formulation of the SP algorithm that helps to make clear how individuals with different information are treated in the algorithm. As seen in the proof located in Appendix C, the transformation from one formulation to the other is mechanical and does not rely on any assumptions regarding the signals received by forecasters and their votes or vote meta-predictions.

LEMMA 1. *The SP algorithm can be rewritten as*

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} W_i^{SP} V_i > 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where each forecaster's weight is given by the normalised absolute difference between that forecaster's vote and their vote meta-predictions:

$$W_i^{SP} := \frac{|V_i - M_i^V|}{\sum_{j=1}^{N} |V_j - M_j^V|}.$$

Proof: All proofs are collected in Appendix C.

In the weighted average formulation of the SP algorithm, the weights are constructed so that $\sum_{i=1}^{N} W_i^{SP} = 1$. Thus, the weight given to each individual forecaster is proportional to $|V_i - M_i^V|$, the absolute difference between the forecaster's vote and the forecaster's meta-prediction about the votes of others.

The alternative SC algorithm uses probabilities and probability meta-predictions to predict the true outcome. Analogous to the SP algorithm, the average probabilistic forecast (or confidence) is compared to the mean probability meta-prediction. If the mean probabilistic forecast is larger than the mean probabilistic meta-prediction, the event is predicted to be true. Otherwise, the event is predicted to be false. Formally,

$$T_{SC}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N}(P_i - M_i^P) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Like the SP algorithm, the SC algorithm can be represented as a weighted average. In this representation

$$T_{SC}(X) = \begin{cases} 1 & \textit{if } \sum_{i=1}^{N} W_i^{SC} \mathbb{I}_{\{P_i > M_i^P\}} > 0.5 \\ 0 & \textit{otherwise,} \end{cases}$$

where (i) $\mathbb{I}_{\{P_i > M_i^P\}}$ is an indicator variable that is one when a forecaster's probability forecast exceeds their probability meta-prediction and zero otherwise and (ii) each forecaster's weight is given by the normalised absolute difference between the forecaster's probabilistic forecast and their probability meta-prediction:

$$W_i^{SC} := \frac{|P_i - M_i^P|}{\sum_{j=1}^{N} |P_j - M_j^P|}.$$

The weighted version of the SC algorithm has the same structure as the weighted version of the SP algorithm, but has two differences. First, the algorithm uses the difference between a forecaster's probabilistic forecast (or confidence) and their probability meta-prediction (rather than their vote and vote meta-prediction) to identify whether a forecaster should be recorded as a zero or a one in the final aggregation. As discussed below, an individual who receives $\sigma_i > s_\varnothing$ is predicted to have a probability forecast that exceeds their probability meta-prediction while the opposite is true when $\sigma_i < s_\varnothing$. Thus, the algorithm assigns a forecaster the equivalent of a true vote when they have a signal greater than the prior, and a false vote when they have a signal less than the prior. Second, the SC uses the probability forecasts and probability meta-predictions to generate the weights rather than using the votes to generate the weights. As discussed below this seemingly small adjustment has important implications in the way that the two algorithms weight forecasters with different signals.

2.2. Weights and Information

We first ask how the weights used in the SP and SC algorithms relate to information when all forecasters reports are consistent with Bayes rule. Intuitively, an algorithm will be able to best exploit the private information of forecasters if forecasters with more informative private signals contribute more to the algorithms final performance than those who have less informative private signals. The following propositions shows that the opposite relationship holds in the SP algorithm in situations where the prior is unbiased:

Proposition 1. *In the SP algorithm, if (i) forecaster i has a more informative private signal than j and (ii) the prior is unbiased, then the weight given to forecaster i will be strictly less than the weight given to forecaster j.*

The intuition for Proposition 1 can be seen in the left side of panel (a) of Figure 1, which plots out the vote function and a typical vote meta-prediction function over all possible posteriors in the

case of a symmetric information service, which has an unbiased prior. As can be seen by looking at the vote function, individuals will vote $V_i = 0$ when $\sigma_i < 0.5$ and $V_i = 1$ when $\sigma_i > 0.5$. Thus, the vote is a step function that switches exactly at the unbiased prior.

The vote meta-prediction of an individual is based on their belief about the votes made by all other participants. Given an outcome state o, the expected proportion of true votes from information service Q is given by

$$\mathbb{E}V(Q|o) = \sum_{\{i|s_i \geq .5\}} \gamma(Q_{oi}),$$

where $\gamma(Q_{oi}) = \frac{1}{2}Q_{oi}$ if $s_i = .5$ and $\gamma(Q_{oi}) = Q_{oi}$ otherwise. A forecaster with signal s_k's vote meta-prediction about the average vote share from information service Q is

$$M^V(Q|s_k) = s_k\mathbb{E}V(Q|T) + (1 - s_k)\mathbb{E}V(Q|F).$$

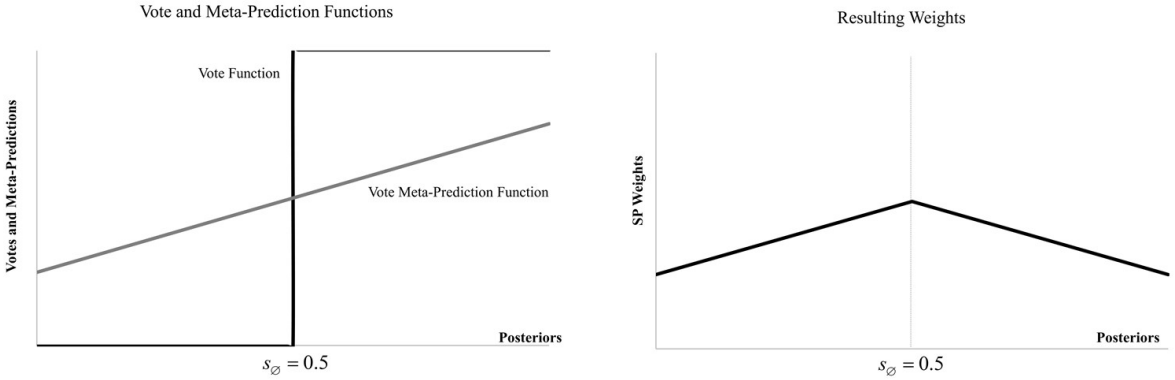
Noting that $\mathbb{E}V(Q|T) > \mathbb{E}V(Q|F)$ when the information service is responsive, $M^V(Q|s_k)$ is a linear function that is increasing in s_k. The underlying information service in panel (a) is symmetric, which implies that $M^V(Q|s_\varnothing) = 0.5$.

As seen in the right side of panel (a), the weights for each individual forecaster is equal to the absolute distance between the vote function and the vote meta-prediction function. This distance is decreasing as the forecasters signal moves away from the prior in both directions. Thus, individuals who have signals closer to the common prior will always have a larger weight than individuals who have signals that are farther away.

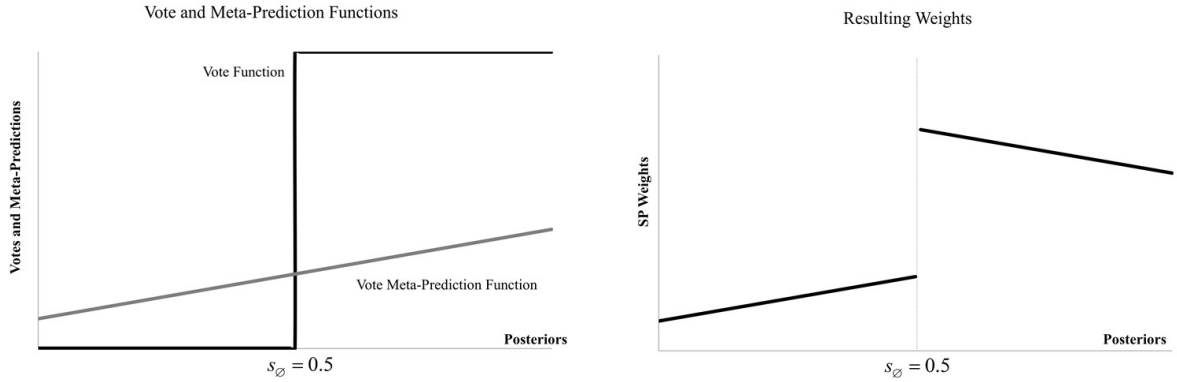
With a symmetric information service, the weighing function is also symmetric and all fully uninformed individual are equally weighted. This is not the case, however, when we consider asymmetric information services. As seen in panel (b) of Figure 1, when the information service is asymmetric, $M^V(Q|s_\varnothing)$ does not necessarily pass through 0.5. As such, there is a gap in the weight function at $s_\varnothing$. This gap is the main way in which the SP algorithm is able to correct for asymmetries in the information service that leads majority voting algorithms to incorrectly predict the state. In particular, an individual who votes for true but predicts that others are more likely to vote false reveals commonly known information about the properties of the information service. This information is then used to increase the weights of individuals who vote against the most popular outcome.

Despite the algorithm taking advantage of information about the asymmetry of the information service, individuals who have signals closer to the common prior always have a larger weight than individuals who have signals that are farther away on the same side of the prior. This implies that forecasters with more informative private signals continue to receive smaller weights than comparable forecasters with less informative private signals.

(a) SP Weights in Symmetric Decision Problems



(b) SP Weights in an Asymmetric Decision Problem



(c) SP Weights in a Biased Decision Problem

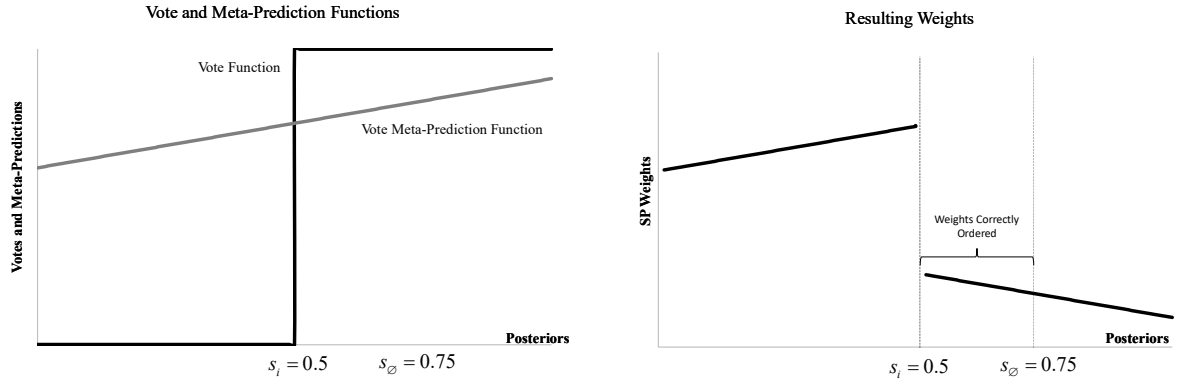


Figure 1 The left panels show the vote function and a typical vote meta-prediction function over all possible posteriors in (a) the case of a symmetric information service with an unbiased prior, (b) an asymmetric information service with an unbiased prior, and (c) a symmetric information service with a biased prior. The right panels show the weights assigned by the SP algorithm for each possible posterior in each of the three cases.

Finally, when the prior is biased, if (i) forecaster i has a more informative private signal than j and (ii) both signals are between the biased prior of $s_\varnothing$ and the uninformed prior of 0.5, then i will be weighted more than j. However, this relationship is reversed in other parts of the distribution. This can be seen in the example shown in panel (c) of Figure 1, where forecasters have a prior of 0.75 and where weights are decreasing for posteriors greater than $s_\varnothing = 0.75$ and for signals that are below 0.5.

We now show that the weights in the SC algorithm is well ordered when it comes to the information contained in the forecaster's private signals:

PROPOSITION 2. *In the SC algorithm, if forecaster i has a more informative private signal than forecaster j, then the weight given to forecaster i will be strictly greater than the weight given to forecaster j.*

The intuition for Proposition 2 can be seen in the left side of panel (a) of Figure 2, which plots out the probability forecast function and a typical probability meta-prediction function over all posteriors in the case of a symmetric information service. As can be seen, the probability forecast function is a linear line with a slope of 1. The probability meta-prediction function is also linear and is based on their belief about the probability of all other participants. Given an outcome state o, the expected average forecast from information service Q is given by

$$\mathbb{E}P(Q|o) = \sum_{s_i} s_i Q_{oi}.$$

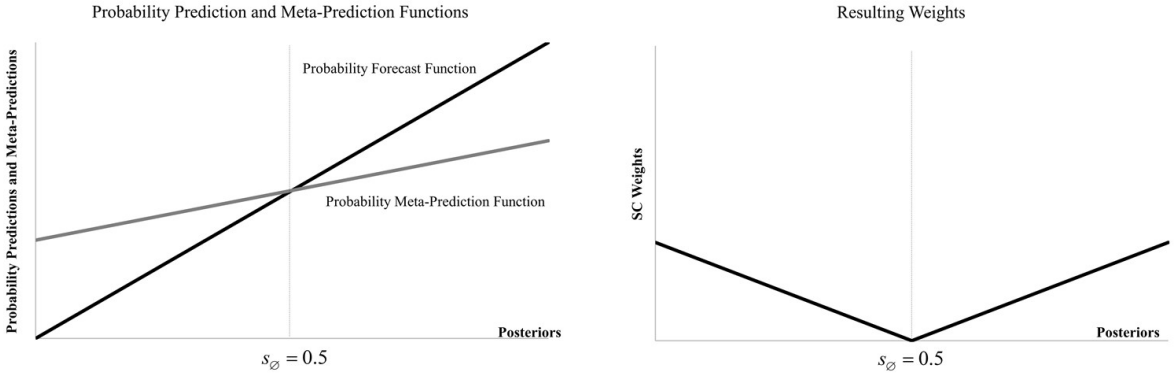
A forecaster with signal s_k's probability meta-prediction about the forecast of others is given by

$$M^P(Q|s_k) = s_k \mathbb{E}P(Q|T) + (1 - s_k)\mathbb{E}P(Q|F).$$

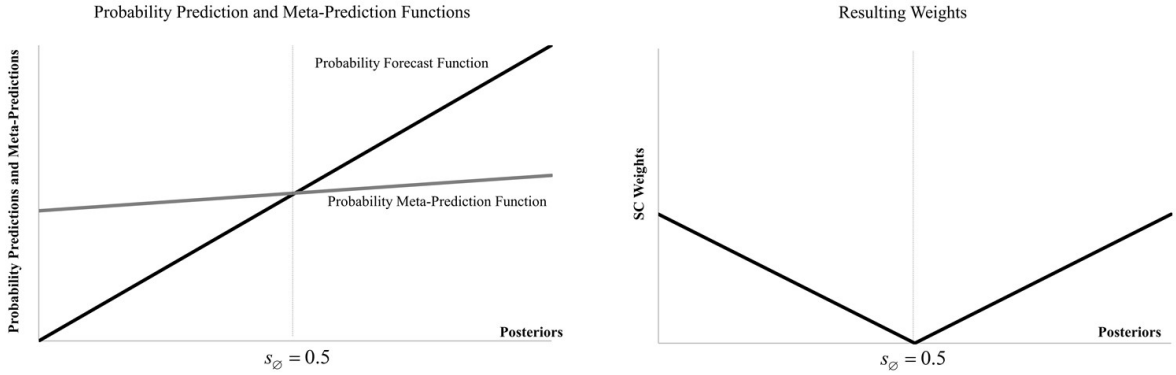
By the law of iterated expectations, $\mathbb{E}P(Q) = s_\varnothing \mathbb{E}P(Q|T) + (1 - s_\varnothing)\mathbb{E}P(Q|F)$. Thus, $\mathbb{E}P(Q|T) > \mathbb{E}P(Q|F)$ and $M^P(Q|s_k)$ is a linear function that is increasing in s_k with a slope less than 1. The law of iterated expectations also implies that the two lines will intersect at the prior of 0.5. The net difference between the two lines generates a "v" shape that correctly orders forecasters in terms of the informativeness of their signals.

Panels (b) and (c) of Figure 2 show that the mechanism also correctly weighs forecasters according to the informativeness of their signals in asymmetric problems and biased problems. As seen in panel (b), asymmetric information services do not substantially change the way the algorithm operates since both probabilities and probability meta-predictions increase linearly in the posterior.

(a) SC Weights in Symmetric Decision Problems



(b) SC Weights in an Asymmetric Decision Problem



(c) SC Weights in a Biased Decision Problem

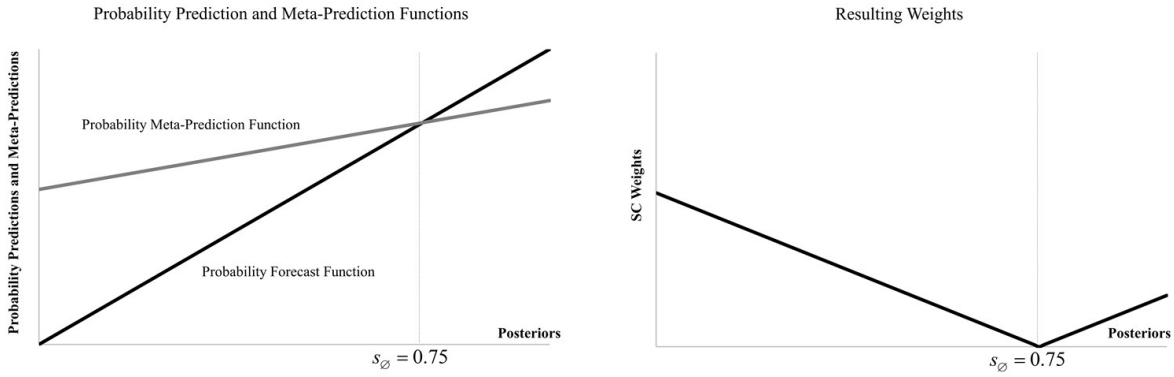


Figure 2 The left panels show a typical probability forecast function and probability meta-prediction function over all possible posteriors in (a) the case of a symmetric information service with an unbiased prior, (b) an asymmetric information service with an unbiased prior, and (c) a symmetric information service with a biased prior. The right panels show the weights assigned by the SC algorithm for each possible posterior in each of the three cases.

As seen in panel (c), in a biased problem, the probability function and meta-probability line cross at the prior. Thus, individuals who receive no signal will still have zero weight.[7]

The different pattern of weights has implications for the accuracy of the SP and SC algorithms. The **expected weight assigned to true in the SP algorithm** as N grows large is

$$\mathbb{E}[W^{SP}] := \lim_{N\to\infty} \sum_{i}^{N} W_i^{SP} V_i.$$

Similarly, the **expected weight assigned to true in the SC algorithm** as N grows large is

$$\mathbb{E}[W^{SC}] := \lim_{N\to\infty} \sum_{i}^{N} W_i^{SC} \mathbb{I}_{\{P_i > M_i^P\}}.$$

The following proposition shows that these expected weights are ordered in unbiased decision problems and that the SC algorithm will always assigns more weight to the correct state as N grows large:

PROPOSITION 3. *For any unbiased information service,* $\mathbb{E}[W^{SC}] \geq \mathbb{E}[W^{SP}]$ *when the correct answer is true and* $\mathbb{E}[W^{SC}] \leq \mathbb{E}[W^{SP}]$ *when the correct answer is false.*

Based on the work by Prelec et al. (2017), $\mathbb{E}[W^{SP}] \geq 0.5$ when the correct answer is true and $\mathbb{E}[W^{SP}] \leq 0.5$ when the correct answer is false. Thus, in very large samples, both the SP and the SC answer will generate the correct answer. In small samples, the sample distribution will converge to a normal distribution with a mean equal to the expected weight assigned to true. This implies that if the variance of the two algorithms are the same, the SC algorithm will be more accurate than the SP algorithm.

In Appendix A, we report results from numerical simulations where we randomly constructed 100,000 unbiased information services and calculated the variance in the total weight assigned to the correct state. We find that the variances of the two algorithms are similar in magnitude. We also calculate the sample size necessary to predict the correct state in 97.5% of cases under the assumption that the distribution is approximately normal in each sample. The SC algorithm requires a smaller sample size than the SP algorithm in over 99% of cases.

Appendix A also includes an analytic example where we explore how the SP and SC algorithms behave in a heterogeneous environment consisting of fully informed forecasters and forecasters who receive only weak signals. We show that in this setting, the SP algorithm may require much larger

[7] We note that the confidence-weighted algorithm, which calculates the average of all probabilistic forecasts and predicts true if this value is above 0.5 and false if the value is below 0.5, can also be written as a weighted average where the numerator of each weight is equal to $|P_i - 0.5|V_i$. This algorithm will generate "v" shaped weights that are centred at 0.5. Thus, in unbiased problems, if forecaster i has a more informative private signal than forecaster j, i will have a larger weight. This relationship does not hold, however, in biased problems because a forecaster with a more informative signal may have a posterior that is closer to 0.5.

samples to ensure a high level of accuracy because the forecasters with weak signals will have large weights. This example suggest that the difference in weight functions may be important in difficult problems where there is only a small fraction of forecasters who know the correct answer. We study how the performance of the two algorithms relate to task difficulty in Section 3.

2.3. The Weighting of Experts

Although our first result suggests that the SP algorithm may over-weight uninformed individuals and under-weight informed ones, it ignores a key advantage of crowd forecasts. In problems with an unbiased prior, the votes of forecasters who receive no information will be random while the votes of those who know the correct state will be perfectly correlated. This will cause the votes of the uninformed forecasters to partially cancel out as crowd size increases and may offset the weighting of individuals.

To understand the aggregate properties of both algorithms, we consider a more general environment in which individuals have access to one of two information services that are ordered in terms of informativeness. We refer to experts as individuals who draw signals from the more informative information service and novices as individuals who receive draws from the less informative service. Thus an expert is defined as individual who is expected to be better informed about the correct answer prior to being asked a particular question.

We consider a variation of our baseline environment where we consider the limiting case where N is countably infinite. We divide forecasters in the population into two groups: experts and novices. Let Q^E be the information service used by expert forecasters and let Q^N be the information service used by novices. We assume that the proportion of experts in the crowd is known to all parties and given by $\theta \in [0,1]$. We also assume that the properties of Q^E and Q^N are common knowledge.

We continue to assume that all forecasters make reports that are consistent with Bayes rule and we make three additional assumptions regarding the information services used by novices and experts.

Assumption 1. *Information service Q^E is more informative than information service Q^N: there exists a non-negative stochastic matrix $Z = [Z_{ki}]_{(m+1)\times(m+1)}$ such that*

$$Q^N = Q^E Z.$$

Assumption 1 says that when Q^E is more informative than Q^N, $Q^N_{oi} = \sum_k Q^E_{ok} Z_{ki}$. As we are multiplying across the rows of Q^E, we can interpret Z_{ki} as the conditional probability that when message k is received by Q^E, message i was received by Q^N. Thus $Z_{ki} = p(s_i|s_k)$ and Q^E is more

informative than Q^N if it is possible to garble the signals of Q^E and generate Q^N.[8] Note that Z is a non-negative stochastic matrix with $\sum_i Z_{ki} = 1$.

ASSUMPTION 2. *Experts and Novices draw conditionally independent signals: for a signal s_i from Q^N and a signal s_k from Q^E,*

$$p(s_i, s_k) = p(s_i|T)p(s_k|T)p(T) + p(s_i|F)p(s_k|F)p(F).$$

ASSUMPTION 3. *Information service Q^E is responsive.*

Assumptions 2 extends the assumption that signals are conditionally independent after conditioning on the state to an environment with two information services. The assumption rules out perverse situations where the garbling matrix creates additional information about the signals of others. Assumption 3 requires that at least expert forecasters will vote for the correct state with a positive probability. This assumption is necessary for the SP algorithm because it is vote based, but is not required for any result related to the SC algorithm.

Assumptions 1 and 2 imply that the information services are ranked but that signals from the two information services are independent once we condition for the state. Assumption 2 is sufficient for the monotone likelihood ratio property (MLRP) to hold for signals between any two information services. This property implies that when an individual receives a high signal, he believes that other forecasters are also more likely to receive a high signal.

LEMMA 2. *For signals $s_i > s_j$ drawn from Q^t, $t \in \{N, E\}$, and signals $s_k > s_l$ drawn from Q^τ, $\tau \in \{N, E\}$, the monotone likelihood ratio property holds:*

$$p(s_i|s_k)p(s_j|s_l) > p(s_j|s_k)p(s_i|s_l). \tag{1}$$

Assumption 3 ensures that when the prior is biased, a subset of experts are willing to change their vote away from the prior for at least some realisation of the signal. Combined with MLRP, this assumption is enough to prove a modified version of PSM's theorem regarding the average estimates of the votes:

LEMMA 3. *In the SP algorithm, if Assumptions 1-3 hold, then the average estimate of the votes for the correct answer will underestimate the true proportion of votes for the correct answer as $N \to \infty$.*

[8] In cases where the signal space is continuous, it is common to allow for an infinite number of forecaster types but to assign each type a simpler two-signal information service with a high and a low signal and where only the columns representing these signals are included. In this two-signal case, Q^A is more informative than Q^B if the posteriors of Q^A "bracket" the posteriors of Q^B. That is, the posteriors that result from receiving the low and high signal from information service Q^A are closer to zero and one, respectively, than the posteriors that result from receiving the low and high signal from information service Q^B. It is also possible to order information services in the continuous case by using Blackwell's Theorem (Blackwell 1951). See Blackwell and Girshick (1979) for a detailed discussion of Blackwell's Theorem and some additional approaches to ordering information services.

The SP mechanism will predict the correct answer if the vote meta-prediction underestimates the true proportion of votes for the correct answer. Thus, Lemma 3 implies that the SP mechanism will continue to predict the correct answer in the limit when there are both experts and novices. The following lemma shows that the SC algorithm has a similar property when Assumption 2 holds:

LEMMA 4. *In the SC algorithm, if Assumptions 1-2 hold, then the average probability meta-prediction will be below the average probability forecast when the state is true and above the average probability forecast when the state is false as* $N \to \infty$.

Note that when $Q^E = Q^N = Q$, Assumptions 1 and 2 always hold. Thus Lemma 4 implies that the SC algorithm is robust to bias in the initial model where all forecasters draw signals from the same information service.

2.4. The Expected Total Contribution of an Expert or Novice

We now turn to the question of how the SP and SC algorithms weight experts and novices. Given information services Q^E and Q^N, a forecaster with signal s_k will make a vote meta-prediction of

$$M^V(\theta|s_k) := \theta M^V(Q^E|s_k) + (1-\theta)M^V(Q^N|s_k),$$

where θ is the proportion of experts in the environment, $M^V(Q^E|s_k)$ is the vote meta-prediction of forecasters from information service Q^E, and $M^V(Q^N|s_k)$ is the vote meta-prediction of forecasters from information service Q^N. Thus, the expected vote meta-prediction of forecasters in information service Q^t $(t \in \{N, E\})$ when the state is o is given by

$$\mathbb{E}[M^V(\theta|Q^t, o)] := \sum_k M^V(\theta|s_k)Q^t_{ok}.$$

Likewise, a forecaster with signal s_k will make a probability meta-prediction of

$$M^P(\theta|s_k) := \theta M^P(Q^E|s_k) + (1-\theta)M^P(Q^N|s_k).$$

Thus, the expected probabilistic meta-prediction in information service Q^t $(t \in \{N, E\})$ when the state is o is given by

$$\mathbb{E}[M^P(\theta|Q^t, o)] := \sum_k M^P(\theta|s_k)Q^t_{ok}.$$

The quantities $\left[\mathbb{E}[V(Q^t|T)] - \mathbb{E}[M^V(\theta|Q^t, T)]\right]$ and $\left[\mathbb{E}[M^V(\theta|Q^t, F)] - \mathbb{E}[V(Q^t|F)]\right]$ represent the expected difference between the votes of forecasters with information service Q^t and their vote meta-prediction, for the true and false states respectively. We will call these quantities the **expected total contribution of an expert or novice** in the SP algorithm in state T and F respectively since they represent the total expected impact of a randomly selected individual from a

given group taking into consideration both their vote and their vote meta-prediction under the given state. Similarly, we will call $\big[\mathbb{E}[P(Q^t|T)] - \mathbb{E}[M^P(\theta|Q^t,T)]\big]$ and $\big[\mathbb{E}[M^P(\theta|Q^t,F)] - \mathbb{E}[P(Q^t|F)]\big]$ the expected total contribution of an expert or novice in the SC algorithm.

In state T, the expected total contribution of an expert exceeds the expected contribution of a novice in the SP algorithm if

$$\big[\mathbb{E}[V(Q^E|T)] - \mathbb{E}[M^V(\theta|Q^E,T)]\big] > \big[\mathbb{E}[V(Q^N|T)] - \mathbb{E}[M^V(\theta|Q^N,T)]\big].$$

This leads us to our definition for leveraging expertise:

Definition 5. An algorithm **leverages expertise** if the expected total contribution of an expert exceeds the expected contribution of a novice in all states.

In the SC algorithm, an individual's weight is strictly increasing in their signal. Using Blackwell's theorem, we can show the following result:

Proposition 4. *The SC algorithm leverages expertise in all environments where Assumptions 1 and 2 hold.*

Proposition 4 shows that in a very general set of decision problems, the SC algorithm is able to leverage expertise. The result naturally generalizes to any number of information services as long as they are ranked in terms of informativeness. Thus, under a wide range of problems, the mechanism is predicted to leverage expertise.

In contrast, in Appendix B we provide two counter examples where the SP algorithm fails to leverage experts. These examples show that when the information services is asymmetric, it is possible to find information services where the total contribution of experts is less than that of novices in at least one state. Thus, Assumptions 1-3 are not sufficient to ensure that experts are leveraged and the SP algorithm may be less effective in heterogeneous environments. We provide two additional properties of the information service that are sufficient to ensure that the SP algorithm leverages expertise in symmetric decision problems and provide an example that helps to explain where these additional properties come from. The example suggests that the SP algorithm is likely to perform best in problems where there is a moderate number of experts.

2.5. Properties of the SC algorithm when confidence measures are noisy

Thus far we have considered how the SC algorithm behaves in an ideal setting where all forecasters are Bayesian and where the confidence elicited by each individual coincides with their posteriors. In this section we discuss some strengths and weaknesses of the SC mechanism that arise when we move from the this ideal setting to one where we incorporate the known biases that exist when eliciting confidences.

As noted in the introduction, a key issue for confidence-weighted algorithms is that they are sensitive to particular types of overconfidence. For instance, as discussed by Hertwig (2012), using confidences to weight forecasters can be problematic in environments where individuals who hold the majority opinion are overconfident while individuals who hold the minority opinions are underconfident. Such environments can arise when confidences are correlated with the majority viewpoint rather than perfectly relating to accuracy (Koriat 2008).

A surprising result is that in large samples, the SC algorithm will continue to correctly predict the correct state in settings where overconfidence occurs under two assumptions about confidences and probability meta-predictions. First, on average, confidences must be increasing in the underlying posterior of an individual forecaster. Second, forecasters must incorporate both their own overconfidence and the overconfidence of others into their meta prediction. We discuss this assumption below after a formal description of our result.

We begin by generalizing the model of Section 2.3 to allow for errors in the relationship between signals and reported confidences.

Definition 6. Forecasters are **systematically miscalibrated** if there exists a weakly increasing function $c\colon [0,1] \to [0,1]$ and a right stochastic matrix $[R_{ij}]_{m+1,m+1}$ such that (i) the probability that an individual with posterior s_i reports confidence $c(s_j)$ is given by $R_{ij} := p(c(s_j)|s_i)$; (ii) for any two posteriors $s_i > s_j$, $c(s_i) \geq c(s_j)$; (iii) there exists two posteriors $s_i > s_j$ that occur with positive probability where $c(s_i) > c(s_j)$; and (iv) $\sum_j c(s_j)R_{ij} = c(s_i)$ for all i.

Our definition of systematic overconfidence allows for forecasters to systematically misapply Bayes rule and to report confidences that are both too high and too low relative to the true posterior. The confidence function, $c(\cdot)$, allows for almost any non-decreasing mapping from true posteriors to confidence reports while the additional error structure allows for additional noise between signals and reports. This error structure is very general and can facilitate most behavioural patterns of overconfidence observed in the literature. In particular, it can accommodate the two main behavioural patterns of overconfidence discussed in Liberman and Tversky (1993) and Griffin and Brenner (2004): general overconfidence, the tendency for all forecasters to assign probabilities that are too close to 1 for the choice that they believe is correct; and specific overconfidence, the tendency for forecasters who believe one answer is correct to assign probabilities that are too close to 1 and for forecasters who believe the other answer is correct to assign probabilities that are too

close to 0.5.[9] [10] It can also accommodate patterns of under-confidence, which is sometimes found in decision problems that are easy (Erev et al. 1994).

When forecasters are systematically miscalibrated, the average confidence of individuals from information service $t \in \{N, E\}$ in state $o \in \{T, F\}$ is given by

$$\mathbb{E}C(Q^t|o) = \sum_i \left(\sum_j c(s_j) R_{ij} \right) Q_{oi} = \sum_i c(s_i) Q_{oi}.$$

We will say that a forecaster's probability meta-prediction is **fully adaptive** if their meta-prediction (i) uses their confidence to assess the likelihood of each state of the world, and (ii) fully predicts the overconfidence of both novices and experts. Thus, an individual who is fully adaptive would report that the average confidence for forecasters from information service Q^t is:

$$M^C(Q^t|c(s_k)) = c(s_k)\mathbb{E}C(Q^t|T) + (1 - c(s_k))\mathbb{E}C(Q^t|F).$$

The following proposition shows that under the assumption of fully adaptive meta-predicitons, the SC algorithm will generate the correct answer for any decision problem where confidence reports are systematically miscalibrated:

PROPOSITION 5. *If forecasters are systematically miscalibrated and all forecasters have fully adaptive meta-predictions, then the average probability meta-prediction will be below the average reported confidence when the state is true and above the average reported confidence when the state is false as* $N \to \infty$.

Although our theoretical result requires a strong assumption about the average meta-prediction in the population, there are reasons to suspect that the algorithm will improve upon other confidence-weighted algorithms even when the assumption does not hold. As discussed in Koriat (2008), Koriat (2012), and Hertwig (2012), confidence-weighted algorithms typically fail in "wicked" problems where the position held by the consensus is wrong. In these problems, individuals who endorse the consensus answer tend to be over-confident while those who endorse the minority answer tend to be under-confident. For the SC algorithm to improve forecasts relative to the confidence-weighted algorithm, the average probability meta-prediction must be above 0.5 when the majority of forecasters vote 'true' but the correct answer is 'false', and below 0.5 if the majority

[9] Although we allow only mean zero errors to be added at each confidence, the relatively weak conditions imposed on the confidence function, $c(\cdot)$, means that we can also model truncation bias that may occur when confidences have symmetric errors that are truncated on $[0, 1]$. In this case, $c(s_i)$ would simply be equal to the expectation of s_i over all realizations of the error.

[10] In fact, as we show in Appendix E, both types of overconfidence can be observed in our empirical results.

of forecasters vote 'false' but the correct answer is 'true'.[11] This will be the case if the average probability meta-prediction and the consensus answer both lie on the same side of the uninformed prior. This relationship is likely to hold if beliefs about the consensus position not only influences each forecasters' confidence report but also their belief about the confidence reports of others.[12]

We note that if forecasters reported $c(\sigma_i) = 0$ when $\sigma_i < 0.5$, $c(\sigma_i) = 1$ when $\sigma_i > 0.5$, and randomizes between 0 and 1 when $\sigma_i = 0.5$, then the fully adaptive meta-prediction would be to report the vote share. Thus, there exists a systematically miscalibrated decision problem where the reports of forecasters coincides with those elicited in the SP algorithm. This insight implies that in settings where forecasters are severely overconfident, the relative rankings of the two algorithms with respect to expertise and average vote weights may not hold. As such, we highlight some other strengths and weaknesses of the two algorithms before moving to the empirical section of the paper.

A clear advantage of the SP algorithm is that it elicits frequency information rather than probabilistic information from forecasters. Vote meta-predictions have the advantage that the forecasters do not have to estimate the level of overconfidence in the environment when forming their belief. Thus, vote meta-predictions may be more accurate in settings where overconfidence is present.[13] Further, a large literature exists that suggests that frequency information is encoded more naturally in the brain and may be more natural for individuals to express (Hintzman et al. 1982, Gigerenzer 1984, Gigerenzer et al. 1991). Thus, the SP algorithm is likely to have lower cognitive requirements than the SC algorithm.

Relative to the SP algorithm, the SC algorithm provides a larger communication space for providing information about signals and meta-knowledge. In particular, the SC algorithm allows for forecasters to reveal that they are (i) uninformed or (ii) have limited insights into the information held be others and gives these forecasters little weight.

[11] For example, suppose that the consensus answer is true, but the correct state is false. Then, if the average meta-prediction is 0.75, the confidence-weighted algorithm will correctly predict false if the average probability forecast is between $[0, 0.5]$ while the SC algorithm will correctly predict false if the average probability forecast is between $[0, 0.75]$.

[12] Note that in cases where the consensus is correct, the SC algorithm will continue to predict the correct answer in large samples as long as the average probability meta-prediction is closer to the uninformed prior than the one calculated with forecasters who are fully adaptive. Thus, as long as forecasters don't over predict the overconfidence of others, the SC algorithm and confidence-weighted algorithm are likely to both perform well in questions where the consensus is correct.

[13] Tereick (2019) argues that vote meta-predictions may be anchored towards the prior and proposes a self-aggregation algorithm that is more robust to these types of biases. An incentive compatible approach proposed by Baillon et al. (2020) to eliciting meta-predictions using a market-based approach with randomized price offers can also potentially be used to mitigate the biases that occur due to overconfidence and inattention.

3. An Empirical Exploration of the SP and SC algorithms

In this section, we empirically estimate the weights generated in the SP and SC algorithms and study how these algorithms treat experts and novices. We concentrate our analysis on two experiments. The first is a replication of the US states capital dataset of Prelec et al. (2017). As seen in Prelec et al. (2017), forecasters in this dataset use what appears to be a heuristic based on population size to predict whether a city is a capital city in problems where they are uninformed. This heuristic naturally leads to a biased prior and is likely to lead to specific overconfidence — the tendency for forecasters who believe the consensus position is correct to be overconfident and forecasters who believe the consensus position is incorrect to be under-confident. We are interested in this environment since the SP is specifically designed to improve forecasting in biased environments and we would predict that this algorithm will perform well.

The second experiment uses a quiz dataset comprised of 500 problems that vary across five levels of difficulty. As seen in the theory section, the relative weighting of experts and novices is related to the proportion of experts in the environment. As we increase the difficulty of decision problems, we would expect the proportion of experts in the dataset to fall. We are thus interested in the relative performance of the SP and SC algorithm as we move from easy problems to hard ones, and we would predict that the SP algorithm leverages expertise most effectively with an intermediate number of experts.

We note that the actual expertise of individuals in our dataset is not observable and thus our empirical strategy requires us to proxy for expertise by using the track record of forecasters on other problems. This proxy is based on the assumption that expertise is correlated across questions and uses the fact that an individual who receives signals from a more informative information service will be correct more often than an individual who receives information from a less informative information service on average.

3.1. Experiment 1

Experiment 1 replicates PSM (2017)'s Study 1, which asked true or false questions about the capital cities of US states. For each state, participants were presented with the largest city and asked whether or not it was the state capital. This dataset provided a natural environment to study the mechanisms underlying the SP algorithm's performance in a biased setting since PSM found in their original study that forecasters typically believe that the largest city in a state is the capital when they do not know the true answer. As this heuristic does not often predict whether a city is the state capital, the underlying information service is likely to be biased in favour of answering true. This allows informed individuals to make meta-predictions that differ substantially from their vote and potentially gives informed individuals large weights.

Our replication used a larger sample size than the original PSM study in order to compare the patterns of predictions and meta-predictions made by the best-performing and worst-performing forecasters in the crowd. In line with PSM, we collected forecasters' votes and meta-predictions about the average vote of others. Additionally, in order to compare the responses used by the SP algorithm and the SC algorithm, we also collected each forecaster's forecast of the likelihood that the event is true and their meta-prediction of the average forecast of all other forecasters.

3.1.1. Methods. We conducted the experiment online, with all participants recruited using Amazon Mechanical Turk. In PSM's experiments, forecasters were monetarily incentivised for accurately predicting the outcome as well as accurately predicting the proportion of the crowd endorsing each response. As our experiment was performed online, we removed the financial incentives to reduce the likelihood of participants looking-up the answer. We tested 100 respondents and only respondents inside the US were able to participate. Each survey was administered using Qualtrics, and participants were paid a flat fee of US $2.50 for completing the survey. Participants were asked to answer each question as honestly as they could, and were asked not to cheat (e.g., by looking up any of the problems online). Eleven individuals who reported cheating at the task or had failed to complete the survey were excluded from the analyses, but were still paid. We completed data collection in January 2020 and analyses were conducted on the data of the remaining 89 participants.

The survey consisted of 50 trials (one for each US state, in alphabetical order of state). On each trial, participants were shown the sentence "X is the capital of Y" where X was the most populous city in the state Y. For example, on the first trial, all participants saw the bolded statement "Birmingham is the capital of Alabama." For each statement, participants were asked to answer four questions:

1. Is this statement more likely to be true or false?
2. What percentage of other people do you think thought the bolded statement was true?
3. What is the probability that the statement is true?
4. What is the average probability estimated by the other forecasters?

Forecasters were restricted to probabilities between 0 and 50 on question 3 if they reported that the statement was more likely to be false and between 50 to 100 if they reported that the statement was more likely to be true. Thus all participants were required to provide votes and probability forecasts that were consistent.

3.1.2. Weights in the SP and SC algorithms: Our theoretical model predicts that the SP weights assigned to individuals will decrease linearly as one moves away from the uninformative posterior of 0.5. However, because the states dataset is predicted to have a biased prior, we would

predict that there will be a gap in the weight function at 0.5 and that this gap may lead false votes to be weighted more than true votes. To test for this, we use an individual's probabilistic forecast as a proxy for the forecaster's posterior[14] and estimate a linear weight function of the form

$$W_{ik}^{SP} = \alpha + \beta_1 |P_{ik} - 0.5| + \beta_2 V_{ik} + \epsilon_{ik}, \tag{2}$$

in which W_{ik} is the numerator of the SP weight of subject i in decision problem k, V_{ik} is their vote, P_{ik} is the probabilistic forecast and ϵ_{ik} are errors that are clustered at the individual level. We use the numerators of the SP weights here as they always fall between 0 and 1 and are fully comparable across problems. We predict that $\beta_1 < 0$, which would indicate that the weights are decreasing in the informativeness of the forecaster's signal between 0 and 0.5 and between 0.5 and 1. Based on PSM, we would also predict that $\beta_2 < 0$, which would indicate that the prior is biased towards true (see panel (c) in Figure 1 for the intuition).

For the SC algorithm, our theoretical model predicts that weights are upward sloping as one moves away from the prior.[15] A proxy for this (unobserved) prior is given by the intersect between the identity line where the probability forecast is equal to itself and a regression line of the probability meta-prediction on the probability forecast. In the states data, this point is at 0.74. We then estimate a linear regression of the form

$$W_{ik}^{SC} = \alpha + \beta_1 |P_{ik} - 0.74| + \epsilon_{ik},$$

in which W_{ik}^{SC} is the numerator of the SC weight of subject i in decision problem k, P_{ik} is the probabilistic forecast and ϵ_{ik} are errors that are clustered at the individual level. We predict that $\beta_1 > 0$, which would indicate that the weights are increasing in the informativeness of signals. We find the following:

Result 1 *Consistent with the theoretical model predictions, weights in the SP algorithm are decreasing in the distance from the 0.5 and there is a large gap in the weight function at 0.5. This gap leads to larger weights for false votes than for true votes. Weights in the SC algorithm are increasing in the distance away from the uninformed prior.*

[14] In a Bayesian framework, an individual's forecast should be their posterior. Although this is not always the case empirically, probabilistic forecasts are strongly predictive of an individuals actual likelihood of being correct in the states dataset. Using a simple linear regression where we regress the probability of being correct on the absolute difference between an individual's probabilistic forecast and the uninformed prior of 0.5, an individual with a probabilistic forecast of 0.5 is correct 46.7 percent of the time while individuals with a probabilistic forecast of either 0 or 1 are correct 65.1 percent of the time.

[15] Note that a biased prior therefore has a qualitatively different effect on the weighting function of the SC algorithm. In the SP weights, a biased prior leads to a gap in the weighting function at 0.5. In the SC weights, a biased prior leads to a shift in the kink point where forecasters are assigned the lowest weight.

Support for Result 1 is given in Figure 3, which plots the relationship between weights and the forecaster's posterior for the SP algorithm (top) and the SC algorithm (bottom). The black solid line in each graph is the predictions from the theoretical models while the dashed line is the estimates from a non-parametric kernel regression.

As seen in the top graph, the magnitude of forecasters' signals ($|P_{i,k} - 0.5|$) is a significant negative predictor of the forecasters' weight in the SP algorithm, $\beta_1 = -0.41$, $F(1, 88) = 40.61$, $p < .001$. Thus, consistent with our predictions, the SP weights appear to be decreasing in the distance away 0.5. Additionally, a forecasters' vote (V_{ik}) is a significant negative predictor of the forecasters weight, $\beta_2 = -0.24$, $F(1, 88) = 73.47$, $p < .001$. This can be seen by the apparent gap in the weight function at 0.5, which suggests a strong bias toward true responses in the dataset.[16] The gap is large enough that the predicted weights of all forecasters voting false are larger than the weights of forecasters voting true in the model specification. As seen below, the gap helps the SP algorithm to predict the correct answer in most of the decision problems.

As seen in the bottom panel, the SC algorithm has weights that are increasing in the distance away from the predicted prior, with a significant and large positive slope in our model that is consistent with our predictions, $\beta_1 = .53$, $F(1, 88) = 69.5$, $p < .001$. On average, better-informed forecasters therefore are generating larger weights. The weights assigned to forecasters who predict that an event is false with certainty are particularly high, with an average weight that is at least twice as large as the weight assigned to any forecaster who voted true.

3.1.3. Expertise in the SP and SC algorithms: Having seen that the weights of our two algorithms match our theoretical predictions, we now explore how forecasters' total contributions relate to expertise. As a first approach, we ranked and sorted forecasters based on their mean accuracy computed using leave-one-out cross-validation and performed a median split between the best-performing individuals ("high-performers") and worst-performing individuals ("low-performers"). For the SP algorithm, we then compared the mean vote for each group to their mean vote meta-predictions for true and false problems separately. For the SC algorithm, we instead compared the mean probability forecast for each group to their mean probability meta-prediction.

The SP and SC algorithms leverage expertise if the average total contribution of an expert exceeds the average total contribution of a novice for both true and false problems. We find the following:

[16] As we show further below, weights in the SC algorithm are increasing in distance away from an uninformed prior of approximately 0.74. This implies that the gap in the SP weighting function is most likely due to a biased prior, rather than forecasters having access to asymmetrical information services. As we see in Appendix D, this also appears to be the case in Experiment 2.

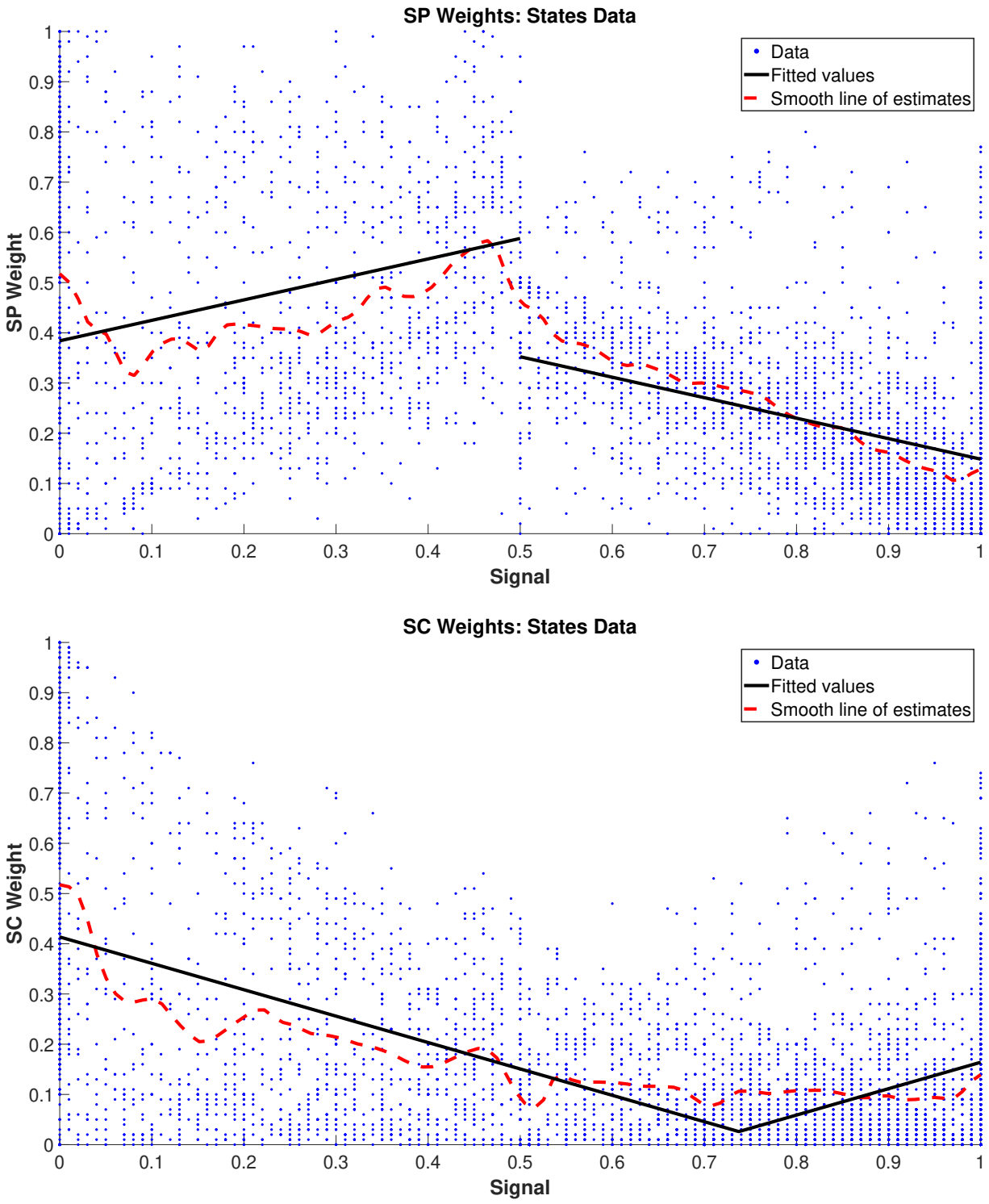


Figure 3 **The relationship between forecasters' posterior and the weight assigned to them by the SP algorithm (top panel) and the SC algorithm (bottom panel) for the States Data. The solid black lines are the predictions from the theoretical models. The dashed line is from a non-parametric kernel regression.**

Result 2 *In the states data, the average total contribution of high-performers in the SP algorithm is statistically greater than that of low-performers in problems that are false, but there is no significant difference in problems that are true. The total contribution of high-performers in the SC algorithm is statistically greater than that of low-performers for both true and false problems.*

Support for Result 2 is provided in Figure 4, which compares the pattern of responses for high-performers and low-performers in the States dataset for both algorithms. The mean of high-performers' responses are shown as red circles, the mean of low-performers' responses are shown as blue crosses, and the shaded regions in these plots indicate where each algorithm would produce correct predictions and where the total contribution of the group has the correct sign. The horizontal (and vertical) distance from the reference line to each point corresponds to the absolute difference between that group's mean vote (or probability forecast) and their mean vote (or probability) meta-prediction. In the top panels, the distance between each point and the dotted line is therefore proportional to the total contribution to the SP algorithm for that particular group and event. Similarly, the distances in the bottom panels are proportional to the total contribution to the SC algorithm for each group and event.

We used paired sample t-tests to compare high-performers' and low-performers' average total contributions separately for problems where the outcome was true and problems where the outcome was false. As seen in the top left panel of Figure 4, high- and low-performers are treated similarly in the SP algorithm for the true problems. The average total contribution of a low-performer was 0.255 while the average total contribution of a high-performer was 0.259. There was no significant difference in high-performers' and low-performers' average total contributions on the 17 true problems in the dataset, $t(16) = 0.506$, $p = 0.62$. On the false problems (the top right panel), the average total contribution of a low-performer was 0.281, while the average total contribution of a high-performer was 0.383. High-performers therefore had significantly higher average total contributions than low-performers on the 33 false problems in the dataset, $t(32) = 9.26$, $p < .001$.

As seen in the bottom set of panels of Figure 4, high-performers have a higher average total contribution in the SC algorithm than low-performers for both true and false problems. On the true problems, the average total contribution of a low-performer was 0.116 while the average total contribution of a high-performer was 0.154. High-performers had a significantly higher total contribution than low-performers on the true problems, $t(16) = 5.35$, $p < .001$. On the false problems, the average total contribution of a low-performer was 0.141 while the average total contribution of a high-performer was 0.247. High-performers therefore also had significantly higher total contribution than low-performers on the false problems, $t(32) = 9.85$, $p < .001$.

In Appendix D, we explore an alternative specification where we divide forecasters into quartiles. Consistent with the results here, forecasters in the best-performing quartile have a higher weight

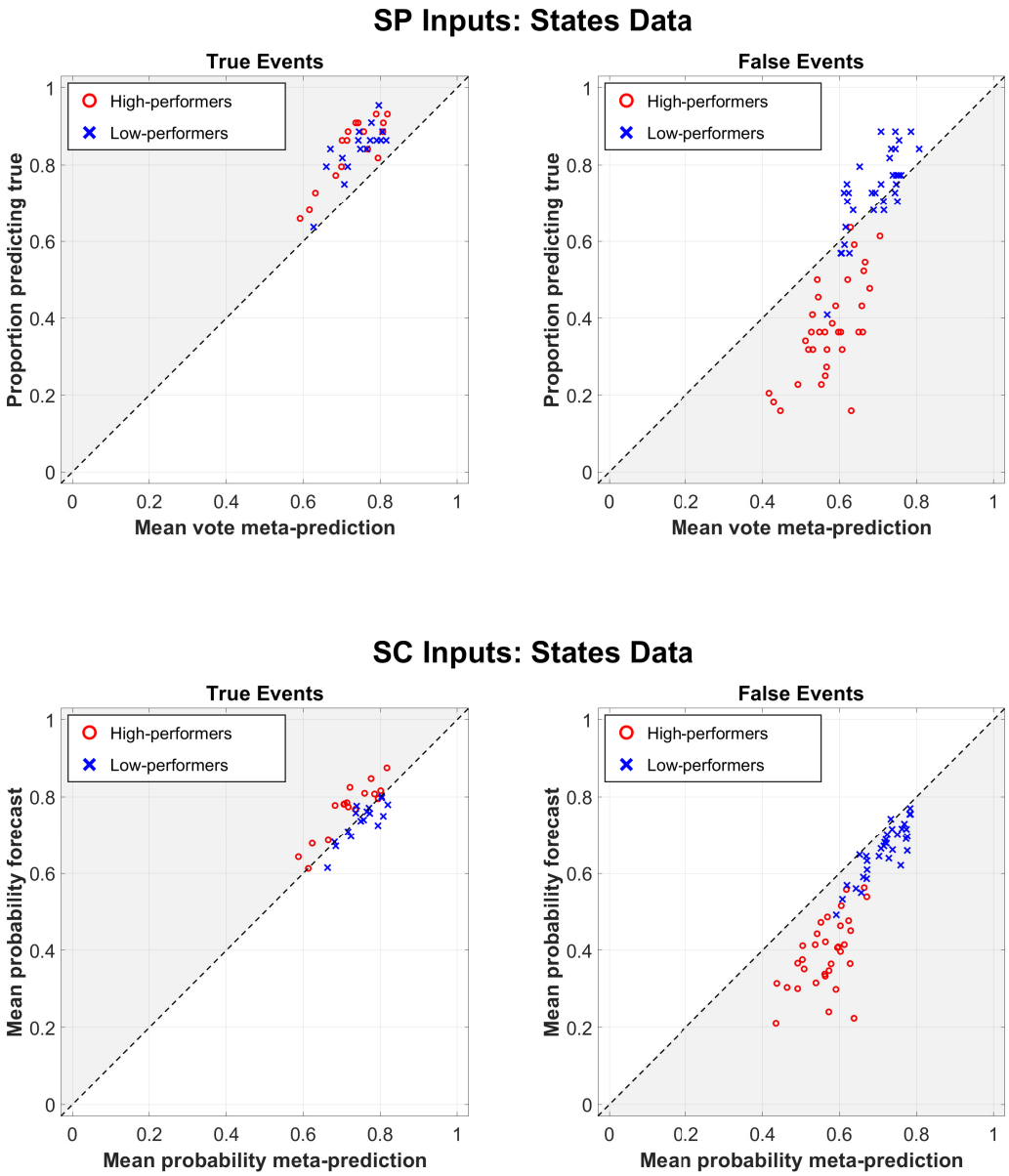


Figure 4 The mean responses from high-performers (red circles) and low-performers (blue crosses) on each question in the States dataset. The top two panels show each group's mean votes compared to their mean vote meta-predictions on the true problems (left) and false problems (right). The bottom two panels show each group's mean probability forecast compared to their mean probability meta-prediction for the true problems (left) and the false problems (right). The diagonal line indicates where each group's vote (or forecast) is identical to their vote (or probability) meta-prediction. The shaded regions indicate where each algorithm would generate correct predictions.

in the SC algorithm than the SP algorithm while forecasters in the worst-performing quartile have a lower weight in the SC algorithm than the SP algorithm.

Taken together, the data from the first experiment supports the results from the theoretical model. The weights in the SP algorithm are decreasing as a participant's probabilistic forecast moves away from the uninformed posterior of 0.5 and the algorithm corrects for bias by generating a discontinuity in the weight function at 0.5. This gap ensures that the total contribution of high-performers exceeds that for low-performers on the false problems, but there is no statistically significant difference for true problems.[17] By contrast, the SC algorithm has weights that are increasing as the probabilistic forecast moves away from the estimated prior. As a result, high-performers are over-weighted by the SC algorithm in both true and false problems.

3.2. Experiment 2

Our theoretical model suggests that the performance of the SP algorithm may vary with the proportion of experts and non-experts in the crowd. To create variation in these proportions, our second experiment explores how the relative performance of the SP algorithm and SC algorithm changes with task difficulty.

3.2.1. Methods. We generated 500 science statements at a US primary and secondary grade school level. Problems were adapted from worksheets on the Education Quizzes website (http://www.educationquizzes.com/us), and then converted into true or false statements. Approximately 2-3 problems were taken from each worksheet from the Biology, Chemistry, Geography, Physics, and General Science categories, spanning from grades 1 to 12, broken up into five levels of difficulty (grades 1 and 2; grades 3, 4, and 5; grades 6, 7, and 8; grades 9 and 10; and grades 11 and 12). We coded "difficulty 1" as the easiest difficulty, and "difficulty 5" as the hardest difficulty. We treated each set of 100 problems of the same difficulty as an individual dataset.

We recruited 500 respondents from Amazon Mechanical Turk; only respondents inside the US were able to participate in the experiment. Participants were paid a flat fee of $4.00 for completing the survey. The survey was conducted on the Qualtrics platform. Participants were asked to answer each question as honestly as they could, and were asked not to cheat (e.g., by looking up any of the problems online). There were 41 individuals who had reported cheating at the task or had failed to complete the survey. These people were excluded from the analyses and analyses were conducted on the data of the remaining 459 participants.

Participants completed 100 trials each, with each trial comprising one statement which was either true or false, and then followed by the same questions we asked in Experiment 1. Half the statements

[17] This result makes sense in light of the weights shown in Figure 3. As seen there, the prior is 0.74 and a forecaster with no information will have a weight that is smaller than a forecaster who knows with certainty that the answer is false but larger than a forecaster who knows with certainty that the answer is true.

at each level of difficulty were true, and the other half were false. Each participant saw 20 statements from each level of difficulty, and statements were presented in one of five randomised orders. Participants who took part in any of our previous experiments were excluded from participating. Data collection for all five datasets was completed in July 2019.

Unlike Experiment 1, we did not force participants' probabilistic forecast to match their votes. Instead, participants who provided votes that were inconsistent with their probability forecasts (i.e., voting "true" but predicting a probability of less than 50% of the statement being true, or voting "false" but predicting a probability of greater than 50% of the statement being true) were excluded from the analysis for that particular question. Approximately 11.3% of responses in the dataset were excluded for this reason.[18]

In Appendix D, we show that the shape of the weight functions and the relative weighting of the algorithms is similar to the results in Experiment 1. Here, we concentrate on how the two algorithms treat experts. We again ranked and sorted forecasters based on their mean accuracy computed using leave-one-out cross validation. We performed a median split between the best-performing individuals and the worst-performing individuals. This exercise was performed for each grades dataset separately and for all five datasets combined. In the analysis for individual grades, mean accuracy was computed using data only from the particular grade. We then computed and plotted the average contribution of high-performers and low-performers for the SP and SC algorithm on the test problems. We find the following:

Result 3 *In the quiz data, the average total contribution of a high-performer is statistically significantly greater than that of a low-performer in both the SP and SC algorithms for both true and false problems.*

Figure 5 shows the average total contribution of high-performers for each algorithm on each dataset. Aggregating across all 500 problems in the dataset, high-performers had a larger average total contribution than low-performers in both the SP algorithm and SC algorithm. For the SP algorithm, low-performers had an average total contribution of 0.228 whereas high-performers had an average total contribution of 0.272; the difference between average total contributions of high-performers and low-performers is significant, $t(499) = 11.6$, $p < .001$. For the SC algorithm, low-performers had an average total contribution to the SC algorithm of 0.098 whereas high-performers had an average total contribution of 0.134. The difference between average total contributions of high-performers and low-performers is also significant, $t(499) = 18.4$, $p < .001$.

[18] We planned to remove inconsistent forecasters prior to running the experiment. However, as the proportion of omitted decisions is relatively large, we also checked to see how both algorithms behave in the full sample. The only substantive difference is noted in footnote 20 below.

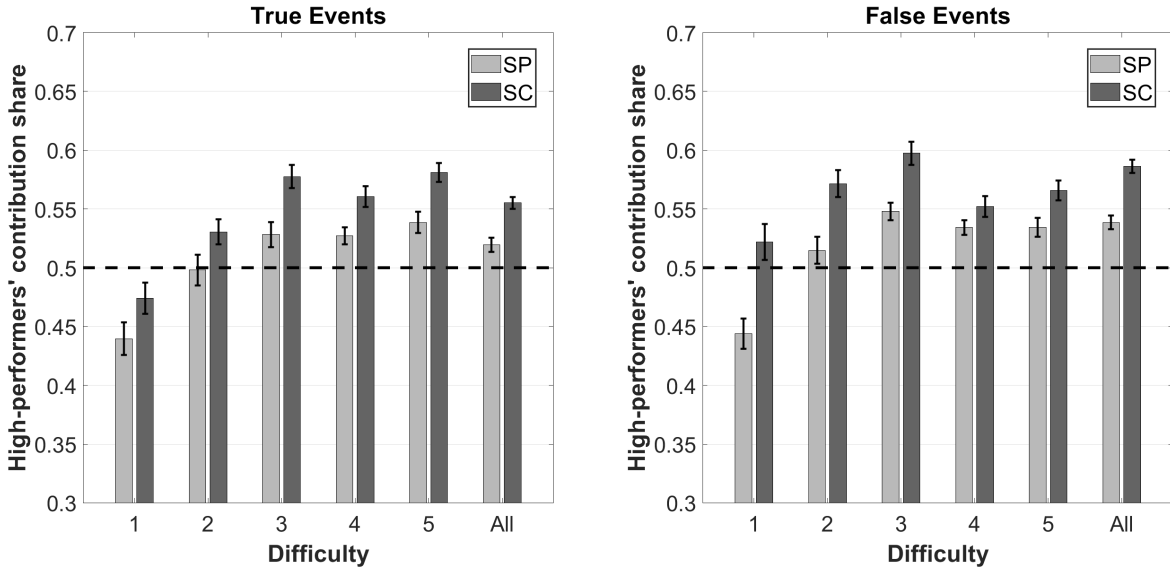


Figure 5 High-performers' average contributions to the SP algorithm and SC algorithm for each of the five individual difficulties and overall across all five difficulties in the quiz dataset. The left panel shows high-performers' share of the crowd contribution on the true events and the right panel shows high-performers' share of the crowd contribution on the false events. The dotted line indicates where high-performers and low-performers have equal contributions to each algorithm's decision.

At the dataset level, high-performers had higher contributions than low-performers in the SP algorithm on all but the easiest difficulty. The SC also assigned greater weights to high-performers than low-performers in all but the easiest difficulty. However, we can see that at both the dataset level and the aggregate level, high-performers' contributions to the SC algorithm (relative to low-performers' contributions) were larger than their contributions to the SP algorithm.

3.3. Performance of the SP and SC algorithms

Thus far we have seen that the SC algorithm leverages informed forecasters both theoretically and empirically and assigns larger weights to forecasters who are correct most often and smaller weights to forecasters who are correct least often. In this section, we study whether these properties translate into improved prediction performance.

To assess prediction performance, we use Matthews correlation coefficient (MCC) as our assessment criterion. This criterion takes into account the large number of false problems in the states dataset, but is similar to accuracy in the quiz datasets where the number of true and false problems are equal. In addition to the SP algorithm and majority voting, we also report the performance of two alternative algorithms that use confidences: the traditional "confidence-weighted" algorithm, which calculates the average probability forecast and assigns a prediction of true if this forecast

exceeds 0.5 and a prediction of zero otherwise, and the max-confidence algorithm, which calculates the average half-range confidence of forecasters who predict true and the average half-range confidence of forecasters who predict false and predicts the larger of these two values.[19]

We use 95% Confidence Intervals (CIs) to test whether there was a statistically significant difference in performance between the SC and the other algorithms. We compute 95% CIs for the mean difference in MCC between the SP algorithm and the SC algorithm for each of the six datasets from Experiment 1 and 2 and in the aggregate over the five quiz datasets. We find the following result:

Result 4 *The SC algorithm has similar performance to the SP algorithm in the States dataset and outperforms the SP algorithm in the Quiz dataset. The performance of the SC algorithm in the Quiz dataset is driven by high accuracy in the more difficult decision problems. By contrast, the SC algorithm performs poorly relative to other algorithms in the easiest decision problems.*

Support for Result 4 is given in Figure 6, which shows the MCC for the SP and SC algorithm relative to the other three algorithms tested. The SC algorithm significantly outperformed the SP algorithm on Grades 3 dataset (95% CI: [.019, .320]) and Grades 4 dataset (95% CI: [.105, .340]). There was no significant difference in performance between the SC algorithm and SP algorithm on States dataset (95% CI: [-.210, .224]), the Grades 2 dataset (95% CI: [-.078, .213]), or the Grades 5 dataset (95% CI: [-.017, .340]). The SP algorithm significantly outperformed the SC algorithm on the Grades 1 dataset (95% CI: [.331, .050]).

In Figure 4, the rightmost set of bars shows the performance of each algorithm after aggregating across all five quiz datasets from Experiment 2. The SC algorithm appears to outperform the SP algorithm, majority vote, and confidence-weighted algorithms by approximately 0.1 in MCC. Computing the paired mean difference in MCC between the SC algorithm and each other algorithm over the five quiz datasets, we find that the SC algorithm had a significantly higher MCC than the SP algorithm (95% CI: [.022, .166]), majority voting (95% CI: [.056, .204]) and the confidence-weighted algorithm (95% CI: [.011, .156]).[20] The SC algorithm also has a higher MCC than the

[19] Although we concentrate on the SP and SC algorithm, it is useful to briefly describe the properties of these alternative mechanisms. Majority voting assigns an equal weight to all forecasters and is guaranteed to leverage experts only in symmetric problems. The confidence-weighted algorithm assigns larger weights to forecasters with more informative private signals in unbiased problems but not biased ones. It is only guaranteed to predict the correct answer in large samples in unbiased forecasting problems and is sensitive to specific overconfidence. Max confidence has the property that the prediction can switch from true to false when the confidence report of one of the forecasters increases. Thus, it isn't possible to represent the algorithm as a weighted average of reports above and below a single posterior threshold. The algorithm is also not guaranteed to leverage experts in any class of decision problems and is not guaranteed to correctly predict the correct answer in large samples.

[20] One caveat to this result is that the exclusion of individuals who were inconsistent appears to negatively impact the performance of the SP algorithm more than the SC algorithm. When inconsistent forecasters are included, the SC algorithm still generally outperforms the SP algorithm and the patterns described above remain. However, the overall difference between the SC and SP algorithm is not statistically significant (95% CI: [-.066, .075]).

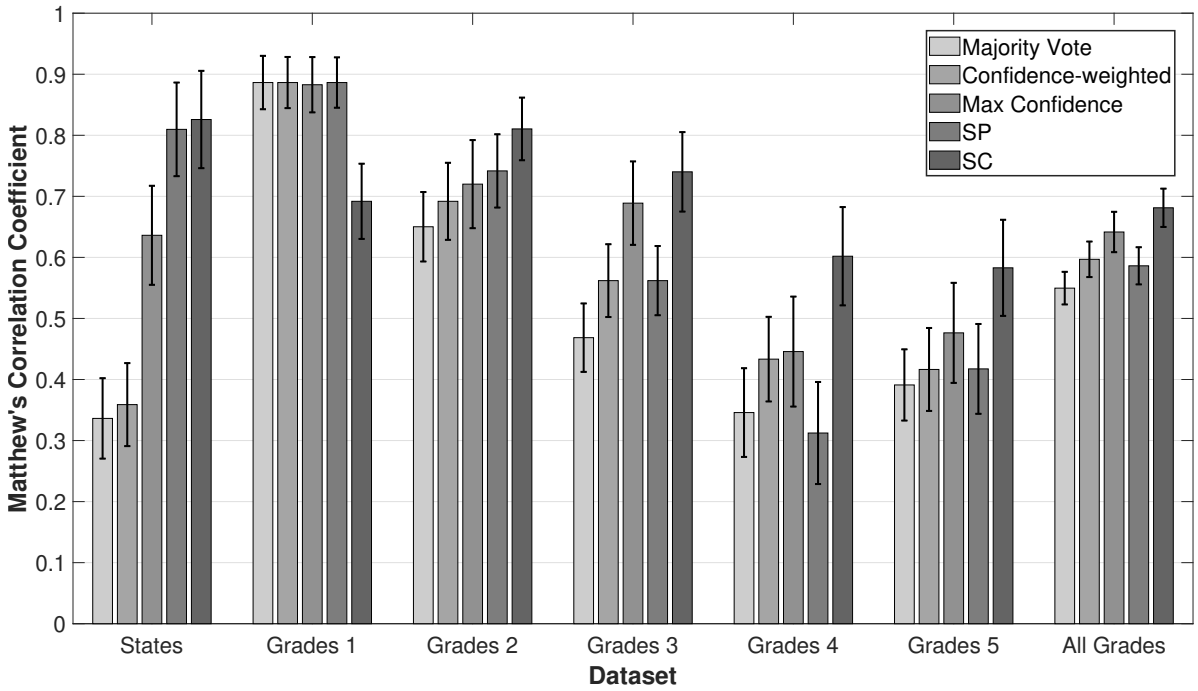


Figure 6 Classification performance of algorithms measured by percentage accuracy on each dataset from Experiment 1 and 2. Error bars show standard error.

max confidence algorithm across all five grade levels, but the difference is not significant (95% CI: [-.039, .120]).

The superior performance on the SC algorithm in hard problems makes sense in the context of the weight functions. In hard problems, many forecasters will have posteriors that are close to the uninformed posterior of 0.5. These forecasters will have large weights in the SP algorithm and this is likely to crowd out the signal from the small number of experts who are likely to exist when the problem is difficult. By contrast, in the SC algorithm, uninformed forecasters will have a zero weight when the prior is unbiased while the small number of expert forecasters in the crowd are likely to have large weights due to having better information.

The relatively poor performance of the SC algorithm in the Grades 1 dataset suggests that the algorithm may not perform well in very easy problems where almost all participants correctly predict the correct answer. We believe this is due to the way that the algorithm handles biased priors. By construction, the SC algorithm eliminates common knowledge from the algorithm by setting the weight of an individual who receives no private signal to zero. If the common knowledge is indeed informative and there is little additional private information, the algorithm may perform poorly. In contrast, the SP algorithm assigns large weights to forecasters who have little private information, and therefore does not eliminate forecasters with common information.

Although it is not the focus of this paper, our theoretical results show that both the SP and SC algorithm are able to correct for bias in large samples and thus any hybrid algorithm that selects

between them based on a secondary criterion will also correctly predict the correct answer. Our data suggests that the SC algorithm performs well on hard problems while the SP algorithm does well on problems that are easy. In principle, an algorithm that switches between these two algorithms based on task difficulty may do better than either algorithm alone. For instance, across our states and quiz datasets, an algorithm that uses the SC algorithm's prediction when the average probability meta-prediction is between $[0.3, 0.7]$ and the SP algorithm's prediction in other circumstances has a MCC of .72 in the quiz data, which significantly outperforms both the SC algorithm (95% CI: [.003, .077]) and the SP algorithm (95% CI: [.066, .200]). It also outperforms both the SP and SC algorithm in the states data, but the difference is not statistically significant (95% CI: [-.057, .257] and [-.070, .259], respectively).

4. Conclusion

Modern forecasting algorithms use the Wisdom of Crowds to produce forecasts better than those of the best identifiable expert. However, these algorithms may be inaccurate when crowds are systematically biased or when expertise varies substantially across forecasters. Recent work by Prelec et al. (2017) has shown that meta-predictions—a forecast of the average forecast of others—can be used to correct for biases even when no external information such as a forecasters past performance is available. Our paper explored how meta-predictions can also be used to improve predictions by identifying and leveraging expertise in the crowd.

We began by outlining an alternative confidence-based version of the SP algorithm. This algorithm retains the theoretical property that it will always predict the correct answers in large samples even when forecasters have a biased prior. In contrast to the SP algorithm, we showed that the SC algorithm weights individuals with more informative private signals more than those with less informative private signals. The algorithm also leverages expertise and can mitigate biases in confidences that arise when individuals who believe the consensus position is correct are overconfident and individuals who believe the consensus position is incorrect are under-confident. Over two experiments, we find that the new SC algorithm does a better job in weighting better-informed forecasters than the original algorithm and show that individuals with higher mean accuracy contribute more to the algorithm than other forecasters.

We also explored the properties of the SP and SC algorithm across a range of problems that varied in difficulty. Overall, the SC algorithm was more effective at leveraging expertise than the SP algorithm. However, the efficacy of the weights did not translate into improved performance at all levels of difficulty. On the easiest problems, the SC algorithm was significantly worse than the SP algorithm, despite the SC algorithm leveraging expertise more effectively. In contrast, the SC algorithm was generally more effective than the SP algorithm on the moderate-to-hard problems.

Thus, despite the theoretical advantages of the SC algorithm, the empirical performance of these algorithms suggests they may be suited to different types of problems, rather than being strictly better or worse than one another.

Overall, our theoretical and empirical findings provide useful insight into how these algorithms can be used to leverage expertise in the single-question domain. The weights used by the SC algorithm have useful properties relating to forecasters' expertise, but importantly, the properties of these weights are not fundamentally tied to each algorithm. Thus, the weights of the SC algorithm can be used independently, for example, for the purposes of improving forecasts in the probabilistic domain (Martinie et al. 2020), or for other purposes such as identifying high-performing individuals for the purposes of compensation or evaluation.

There exist other algorithms that seek to identify expertise in the single question domain, such as those based on forecasters' confidence (Koriat 2008) or decision similarity (Kurvers et al. 2019). These other measures are most effective in 'kind' environments or low-difficulty problems, where the majority of forecasters are likely to vote correctly (Koriat 2008, Kurvers et al. 2019). In contrast, our results suggest that the SC weights are better suited for identifying expertise on moderate-to-high difficulty problems, where the majority of forecasters may often be biased and vote incorrectly. Our results are therefore complementary to the existing literature in that they can be used to identify and leverage expertise in different forecasting environments.

Acknowledgments

We gratefully acknowledge the financial support of the Australian Government Research Training Program (RTP) Scholarship (Martinie), the FBE & MDHS Collaboration Seed Funding Award (Howe and Wilkening), and the Australian Research Council's Future Fellowship Scheme FT190100630 (Wilkening). We also thank Drazen Prelec, H. Sebastian Seung, and John McCoy for providing us access to the data they collected.

References

Baillon A, Tereick B, Wang TV (2020) Follow the money, not the majority: Incentivizing and aggregating expert opinions with Bayesian markets, mimeo.

Blackwell D (1951) Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (The Regents of the University of California).

Blackwell D (1953) Equivalent comparisons of experiments. *The Annals of Mathematical Statistics* 265–272.

Blackwell D, Girshick MA (1979) *Theory of games and statistical decisions* (Courier Corporation).

Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Science* 61(2):267–280.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5(4):559–583.

Condorcet Md Marie Jean Antoine Nicolas de Caritat (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (De l'Imprimerie royale).

Cooke RM (1991) *Experts in uncertainty: Opinion and subjective probability in science* (Oxford University Press on Demand).

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, Johannesson M (2015) Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* 112(50):15343–15347.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological review* 101(3):519.

Fischhoff B, MacGregor D (1982) Subjective confidence in forecasts. *Journal of Forecasting* 1(2):155–172.

Galton F (1907) Vox populi (the wisdom of crowds). *Nature* 75(7):450–451.

Genre V, Kenny G, Meyler A, Timmermann A (2013) Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* 29(1):108–121.

Gigerenzer G (1984) External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology* 97(2):185–195.

Gigerenzer G, Hoffrage U, Kleinbölting (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review* 98(4):506–528.

Gillen B, McKenzie J, Plott CR (2018) Two information aggregation mechanisms for predicting the opening weekend box office revenues of films: Boxoffice prophecy and guess of guesses. *Economic Theory* 65(1):25–54.

Griffin D, Brenner L (2004) Perspectives on probability judgment calibration. Koehler DJ, ed., *Blackwell Handbook of Judgment and Decision*, 177–199 (John Wiley & Sons, Incorporated).

Hertwig R (2012) Tapping into the wisdom of the crowd—with confidence. *Science* 336(6079):303–304.

Hintzman DL, Nozawa G, Irmscher M (1982) Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning and Verbal Behavior* 21(2):128–141.

Koriat A (2008) Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(4):945.

Koriat A (2012) When are two heads better than one and why? *Science* 336(6079):360–362.

Kurvers RH, Herzog SM, Hertwig R, Krause J, Moussaid M, Argenziano G, Zalaudek I, Carney P, Wolf M (2019) How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances* 5(11):eaaw9011.

Lee MD, Lee MN (2017) The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making* 12(4):328.

Liberman V, Tversky A (1993) On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin* 114(1):162–173.

Marschak J, Miyasawa K (1968) Economic comparability of information systems. *International Economic Review* 9(2):137–174.

Marschak J, Radner R (1972) *Economic Theory of Teams (Cowles Foundation Monograph 22)* (Yale University Press, New Haven, CT).

Martinie M, Wilkening T, Howe PDL (2020) Using meta-predictions to identify experts in the crowd when past performance is unknown. *PLoS ONE* 15(4):1–11.

McCoy J, Prelec D (2017) A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint arXiv:1703.04778* .

Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, Baker J, Hou Y, Horowitz M, Ungar L, Tetlock P (2015) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3):267–281.

Müller-Trede J, Choshen-Hillel S, Barneron M, Yaniv I (2017) The wisdom of crowds in matters of taste. *Management Science* 64(4):1779–1803.

Palley A, Satopää V (2020) Boosting the wisdom of crowds within a single judgment problem: Selective averaging based on peer predictions, available at https://ssrn.com/abstract=3504286.

Palley A, Soll JB (2018) Extracting the wisdom of crowds when information is shared. *Management Science* 65(5):1949–2443.

Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532.

Satopää VA, Pemantle R, Ungar LH (2016) Modeling probability forecasts via information diversity. *Journal of the American Statistical Association* 111(516):1623–1633.

Simmons JP, Nelson LD, Galak J, Frederick S (2011) Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research* 38(1):1–15.

Surowiecki J (2005) *The wisdom of crowds* (Anchor).

Tereick B (2019) Improving information aggregation through meta-cognitive judgments, mimeo.

Tetlock PE (2017) *Expert political judgment: How good is it? How can we know?* (Princeton University Press).

Online Appendix A: The relationship between weights and accuracy

In the main text, we provided Proposition 3 that showed that the expected weight of the SP and SC algorithms were ordered. This implies that the means of the limiting distributions are well ordered. We also noted that the variances of the two algorithms are not ordered. To have a better sense of how the two algorithms are likely to perform in small and moderate samples, we construct 100,000 randomly generated unbiased information services using the following process. First, in each state, we draw five uniform $[0,1]$ variables, $x_1,\ldots,x_5$, and set $Q_{oi}=\frac{x_i}{\sum_{i=1}^{5}x_i}$. By construction, the elements of each row sum to one and there will be both a state where $Q_{Ti}>Q_{Fi}$ and a state where $Q_{Ti}<Q_{Fi}$. Thus, each information service will be responsive.

We next generate 1000 samples of size 100 to calculate the mean ($\overline{W}$) and variance ($Var(W)$) of the sample in both the case where the correct answer is true and the case where the sample is false. We use samples of 100 to ensure that the variance generated in re-weighting the observations in each algorithm is taken into account. We also chose this sample size because it is the sample used in our experiments.

Over the $100,000$ samples, the average value of $\overline{W}^{SC}$ is 0.741 and the average value of $\overline{W}^{SP}$ is 0.674 when the state is true. The average observation-level coefficient of variance, $\frac{100Var(W)}{\overline{W}}$, of the SC algorithm is 0.366 while the average observation-level coefficient variance of the SP algorithm is 0.378. Both the difference in means and the difference in the coefficient of variance are significantly different from zero, though the magnitude difference in the coefficient of variances is very small (paired t-test of means: $t(99999)=376.0$, $p<.001$; paired t-test of coefficient of variance: $t(99999)=32.5$, $p<.001$). The results for false question are nearly identical ($1-\overline{W}^{SC}=0.740$ and $1-\overline{W}^{SP}=0.674$ when the state is false; coefficient of variances (using a mean of $1-\overline{W}$ as the denominator) are 0.366 and 0.378 respectively).

Based on the mean and the variance generated by the sample, we approximate the sample size N necessary to ensure an accuracy of 97.5% by finding the point where the lower bound of the confidence interval is equal to 0.5 for both the case where the answer is true and false:

$$\frac{1}{2}=\overline{W}-1.96\frac{(100Var(W))^{\frac{1}{2}}}{N^{\frac{1}{2}}}. \tag{3}$$

The maximum of these two N is the estimated sample size necessary to generate an accuracy of 97.5% for both algorithms. Note that the choice of 97.5% is arbitrary. We chose this threshold because the tail of 2.5% corresponds to the left tail of a two-sided 95% confidence interval. This allows us to use the common multiplier of 1.96 standard deviations when calculating the N in equation (3) above.

Across the $100,000$ samples, the SC algorithm is predicted to require a smaller N in 99.1% of cases while the SP algorithm is predicted to have a smaller sample in only 6 cases. Restricting attention to the 63,280 cases where at least one algorithm requires a sample size of at least 30 and where the central limit theorem is likely to be a reasonable approximation, the SC algorithm is predicted to require a smaller sample size in 99.7% of cases. These results suggest that the SC algorithm is likely to be more efficient in the vast majority of unbiased decision problems in cases where all forecasters are Bayesian.

Figure 7 plots the minimum number of individuals necessary to ensure that the SP and SC algorithms generate the correct forecast 97.5% of the time for two information services containing experts who know

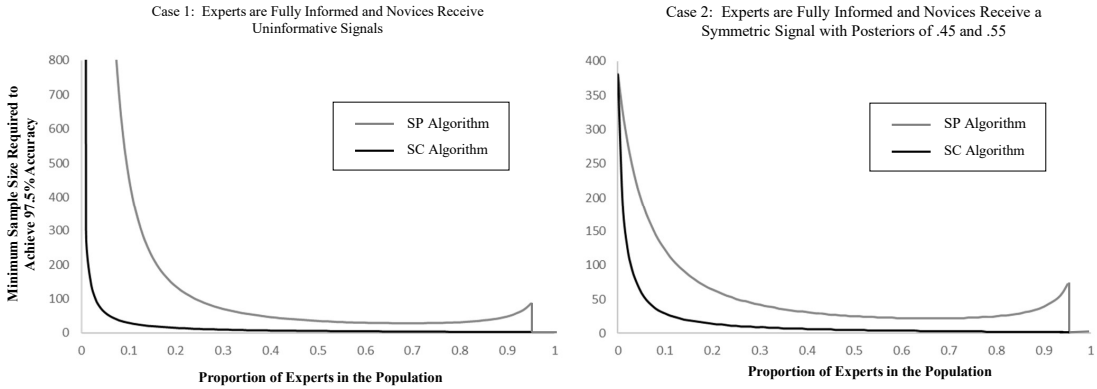


Figure 7 The left panel shows the sample size necessary to achieve 97.5% accuracy with the SP and SC algorithm from a population consisting of fully informed experts and uninformed novices. The right panel shows the sample sizes necessary to achieve 97.5% accuracy with the SP and SC algorithms when novices receive a symmetric signal that generate posteriors of 0.55 and 0.45.

the correct state and uninformed novices. In both panels, forecasters are drawn from a population where a proportion θ are experts and are fully informed about the correct answer and $1-\theta$ are novices. In the left hand panel, novices have no informative signal and both algorithms will be correct 50% of the time when $\theta = 0$ for any N. In the right hand panel, each novice receives an independent signal that is correct 55% of the time and incorrect 45% of the time. In both graphs, we concentrate on a symmetric information service where $s_{\varnothing} = .5$. The cutoffs reported are derived analytically using the exact sample distribution or a normal approximation in cases where the Lindeberg-Lévy Central Limit Theorem applies. We randomly pick a predicted state in cases where either algorithm returns an indeterminate value.

As seen in the left hand panel, the SC algorithm requires a very small sample sizes to accurately predict the correct answer when novices are fully uninformed. This is because uniformed forecasters have zero weight in the algorithm and it only takes a single informed forecaster to generate the correct answer. In the SP algorithm, by contrast, uninformed individuals have a larger weight than the informed forecasters for any θ. Although the expected contribution of each novice is zero, they nonetheless create substantial noise in the algorithm that can lead to inaccurate predictions. As a result, the SP algorithm requires a larger sample than the *SC* algorithm for any proportion of experts. The difference in required sample sizes is particularly pronounced for cases where the proportion of experts is small. For example, when only 10% of the population is an expert, the SC algorithm requires a sample of 29 participants to ensure an accuracy of 97.5% while the SP algorithm requires a sample of 462.

The right hand panel shows that the SP algorithm continues to require larger sample sizes even when the novices are partially informed and that the SC algorithm requires a smaller sample for any proportion of experts. This graph shows that the difference in sample sizes seen in the left hand panel is not due to the assumption that novice forecasters were fully uninformed.

Online Appendix B: Expertise and the SP algorithm

In this appendix, we explore how the SP algorithm treats experts. In part 1 we provide counter examples that show that the SP algorithm does not always leverage experts in a variety of information services. We then provide two additional conditions on the information service that are sufficient to ensure that the algorithm leverages experts in symmetric information services. Finally, we provide an example that highlights some of the intuition that underlines the proof and discuss the cases where we expect the SP algorithm to perform best when forecasters are heterogeneous in their expertise.

Part 1: Counter Examples

Examples 1 and 2 below show that in cases where the prior or the posteriors are asymmetric, it is possible to find counter examples where the expected total contribution of experts in the SP algorithm is less than that of novices in at least one state.

EXAMPLE 1. Consider an environment where $\theta = .5$, the prior $s_\varnothing = .8$ and where the set of additional posteriors are $\{0, .4, .6, 1\}$. Suppose further that the experts' information service over $\{0, .4, .6, .8, 1\}$ is

$$Q^E = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and the Novices' information service is

$$Q^N = \begin{bmatrix} 0 & 0 & .375 & 0 & .625 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Note that the Novices will always vote for True regardless of the state because their posteriors are always greater or equal to .6. Thus, this is a group that is biased and information will only influence their meta-predictions.

We now show that the expected total contribution of expert is *not* greater than the expected total contribution of the novices in the true state. For the experts,

$$\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E, T) = 1 - (.5 * 1 + .5 * 1) = 0,$$

while for the Novices

$$\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T) = 1 - (.5 * .85 + .5 * 1) = .075.$$

Example 1 shows that when the prior is biased, the expected contribution of an expert may be smaller than that of the novice for at least one of the states. The following example shows that even when the prior is unbiased, it is still possible to construct information services where the expected total contribution of an expert is less than that of a novice.

EXAMPLE 2. Consider an environment where $\theta = .5$, the prior $s_\varnothing = .5$ and where the set of additional signal realizations are $\{\frac{x}{x+1}, 1\}$ with $x \in [0, 1)$. Suppose further that the experts' information service over $\{\frac{x}{x+1}, .5, 1\}$ is

$$Q^E = \begin{bmatrix} x & 0 & 1-x \\ 1 & 0 & 0 \end{bmatrix}$$

and the Novices' information service over $\{\frac{x}{x+1}, .5, 1\}$ is

$$Q^N = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

We now show that the expected total contribution of an expert may be lower than the expected total contribution of a novice and that it can be negative for x close to 1. For the experts,

$$\mathbb{E}V(Q^E|T) = 1 - x$$

and

$$\mathbb{E}M^V(\theta|Q^E,T) = .5M^V(Q^E|Q^E,T) + .5M^V(Q^N|Q^E,T) = \frac{1-x}{2(1+x)} + .25$$

In the limit, as $x \to 1$

$$\lim_{x\to 1}[\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E,T)] = 0 - .5^2 = -.25.$$

This is strictly below $\lim_{x\to 1}[\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N,T)] = .25$.

Part 2: Sufficient Conditions

We now discuss two additional properties of the information services that are sufficient to ensure that the SP algorithm leverages expertise in symmetric information services.

Strict Garbling: When information service Q^E is more informative than Q^N, we can find a garbling matrix Z such that each signal in Q^N can be found by adding noise to the signals in Q^E. To guarantee that the SP algorithm leverages expertise, we will require a stronger condition. Rather than using any set of signals from Q^E, we will require that Q^N can be constructed by garbling only signals in Q^E that are at least as informative as the signal being constructed in Q^N. Let $\hat{s} \in [0, 0.5]$ be an arbitrary posterior between 0 and 0.5. Further, let

$$F^t(\hat{s}) := \sum_{\{i|s_i \le \hat{s}\}} [Q^t_{Ti} + Q^t_{T(m+2-i)}]$$

to be the probability of having a posterior that is less than or equal to an arbitrary posterior $\hat{s}$ or greater than or equal to $1 - \hat{s}$ when receiving signals from information service $t \in \{E, N\}$ in state T. Note that in a symmetric information service,

$$\sum_{\{i|s_i \le \hat{s}\}} [Q^t_{Ti} + Q^t_{T(m+2-i)}] = \sum_{\{i|s_i \le \hat{s}\}} [Q^t_{Fi} + Q^t_{F(m+2-i)}]$$

and thus, under symmetry, $F^t(\hat{s})$ is invariant to the state chosen to evaluate it.

Definition 7. Q^N is a **strict garbling** of Q^E if (i) both Q^N and Q^E are symmetric, (ii) $F^N(\hat{s}) \le F^E(\hat{s})$ for all $\hat{s} < 0.5$ and (iii) exists at least one $\hat{s}$ for which $F^N(\hat{s}) < F^E(\hat{s})$.

Problem Difficulty: Forecasting problems are the most difficult when forecasters receive weak signals about the true state and where the vote shares are close to 50:50. We define a forecasting problem as hard if at least a quarter of the population will answer the question incorrectly:

DEFINITION 8. A forecasting problem is **hard** if less than 75% of forecasters vote "true" in the true state and greater than 25% of forecasters vote "true" in the false state.

The following result provides a set of sufficient conditions that ensure that the SP algorithm leverages expertise in environments where Assumptions 1-3 hold and where there are exactly two information services:[21]

PROPOSITION 6. *The SP algorithm leverages expertise if information services Q^N and Q^E are symmetric, Q^N is a strict garbling of Q^E, and the forecasting problem is hard.*

Although the proof of proposition 6 is technical, it is again related to the slope and level of the vote meta-prediction function. We demonstrate this here with a simple example.

Consider a decision problem where the prior $s_\varnothing = .5$ and where the set of additional posteriors are $\{0, .4, .6, 1\}$. Suppose further that the experts' information service over $\{0, .4, .5, .6, 1\}$ is

$$Q^E = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and the Novices' information service over $\{0, .4, .5, .6, 1\}$ is

$$Q^N = \begin{bmatrix} 0 & .4 & 0 & .6 & 0 \\ 0 & .6 & 0 & .4 & 0 \end{bmatrix}$$

In this problem, experts know the correct state, while novices have weak but correctly informative signals. We study how the expected total contributions of novices and experts change with θ.

Since Q^N and Q^E are symmetric, we will restrict attention to the true state. First, note that

$$\begin{aligned} M^V(\theta|s_k) &= \theta M^V(Q^E|s_k) + (1-\theta)M^V(Q^N|s_k) \\ &= \theta s_k + (1-\theta)[.6s_k + .4(1-s_k)]. \end{aligned}$$

For an expert, $\mathbb{E}[V(Q^E|T)] = 1$, and $\mathbb{E}[M^V(\theta|Q^E, T)] = M^V(\theta|s_k = 1) * 1 = \theta + .6(1-\theta)$. Thus

$$\mathbb{E}[V(Q^E|T)] - \mathbb{E}[M^V(\theta|Q^E, T)] = .4 - .4\theta. \tag{4}$$

For the novice, $\mathbb{E}[V(Q^N|T)] = .6$, $M^V(\theta|s_k = .4) = .4\theta + (1-\theta)[.6 * .4 + .4 * .6]$, and $M^V(\theta|s_k = .6) = .6\theta + (1-\theta)[.6^2 + .4^2]$. Thus

$$\begin{aligned} \mathbb{E}[M^V(\theta|Q^N, T)] &= .4M^V(\theta|s_k = .4) + .6M^V(\theta|s_k = .6) \\ &= .4[.4\theta + .48(1-\theta)] + .6[.6\theta + .52(1-\theta)] \\ &= .52\theta + .504(1-\theta). \end{aligned}$$

It follows that

$$\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T) = .6 - [.52\theta + .504(1-\theta)] = .096 - .016\theta. \tag{5}$$

Using equations (4) and (5) above, the total contributions of an expert exceeds that of a novice when $.4 - .4\theta > .096 - .016\theta$ or, equivalently, when $\theta < \frac{19}{24}$.

[21] We note that unlike the proof for the SC algorithm, Proposition 6 does not necessarily generalize to cases where there are more than two information services.

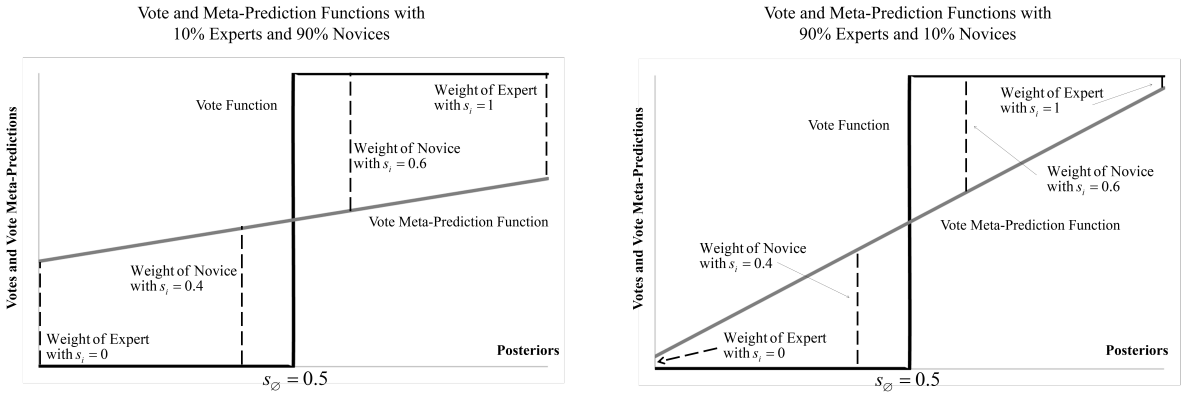


Figure 8 The left panel shows the vote function and vote meta-prediction over all possible posteriors for the case where 10% of the population are experts and 90% are novices. The right panel shows the vote function and vote meta-prediction over all posteriors for the case where 90% of the population are experts and 10% are novices. Weights are equal to the absolute distance between the two functions.

Figure 8 plots the meta-prediction line $M^V(\theta|s_k)$ and the vote function for $\theta = 0.1$ and $\theta = 0.9$. As seen on the left hand side, when $\theta = 0.1$, the meta-prediction line is relatively flat and the weight given to the expert is similar to that of the novices. Because the votes of the novices tend to cancel out while all experts perfectly predict the correct state, experts are leveraged in the decision problem.

By contrast, when $\theta = 0.9$, the slope of the meta-prediction line is close to one and the weights of individuals with high-quality signals grow small. Thus, although some of the forecasts of the novices partially cancel out, the expected contribution of the experts falls below that of the novices.[22]

Within the class of symmetric problems, the SP algorithm is likely to perform best when the total weight given to experts in the algorithm is largest. Our simple example shows that when there are too many experts who know the correct state, the weights given to each individual expert may grow small. Thus, our analysis suggests that the SP algorithm is likely to do best in cases where there are an intermediate number of experts. This is the case with the example above, where the difference between the total expected contribution of all experts, $\theta\big[\mathbb{E}[V(Q^E|T)] - \mathbb{E}[M^V(\theta|Q^E,T)]\big]$, and the total expected contribution of all novices, $(1-\theta)\big[\mathbb{E}[V(Q^N|T)] - \mathbb{E}[M^V(\theta|Q^N,T)]\big]$, is largest at $\theta \approx .616$.

Online Appendix C: Proofs

In this appendix we provide proofs for all the lemmas and propositions in the paper. We provide the proofs to Lemmas (1) – (4) first before presenting the proofs for Propositions (1) – (6).

Lemmas (1) – (4)

Proof of Lemma 1: In this proof, we show that the Surprisingly Popular (SP) algorithm of PSM can be rearranged such that each forecaster's vote is weighted by the normalized, absolute difference between their vote and meta-prediction. We begin with the original form of the SP algorithm and rearrange it to

[22] Note that when $\theta = .9$, 96% of forecasters will vote for the right answer and the problem is not classified as hard. Thus, our sufficient conditions do not cover this case.

show that it is identical to a weighted form, where the weights are given by the absolute difference between their vote and their meta-prediction, normalized by the sum of this difference over all forecasters:

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} \frac{|V_i - M_i^V| V_i}{\sum_{j=1}^{N} |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

In the original SP algorithm, the proportion of the crowd voting for that outcome is compared to the mean meta-prediction, and the most under-predicted outcome is then predicted to be correct. Formally,

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} (V_i - M_i^V) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The crowd for an event with N forecasters can be decomposed into T forecasters who vote true and F forecasters who vote false, $N = T + F$. The report of each forecaster who votes true $t \in \{0, \ldots, T\}$, is given by $X_t := (V_t, P_t, M_t^V, M_t^P)$, and the report of each forecaster who votes false, $f \in \{0, \ldots, F\}$, is given by $X_f := (V_f, P_f, M_f^V, M_f^P)$. The SP equation can therefore be decomposed into

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} (V_t - M_t^V) + \sum_{f=1}^{F} (V_f - M_f^V) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Rearranging this, we get

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} (V_t - M_t^V) > -\sum_{f=1}^{F} (V_f - M_f^V) \\ 0 & \text{otherwise.} \end{cases}$$

As $V_f = 0$, $V_t = 1$, and $M_i^V \in [0, 1]$ the difference between votes and vote meta-predictions for any individual who votes false will always be equal to or less than 0,

$$V_f - M_f^V \leq 0,$$

and the difference between votes and vote meta-predictions for any individual who votes true will always equal or exceed 0,

$$V_t - M_t^V \geq 0.$$

The SP equation is therefore equivalent to

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} |V_t - M_t^V| > \sum_{f=1}^{F} |V_f - M_f^V| \\ 0 & \text{otherwise.} \end{cases}$$

Adding the terms on the left to both sides, we obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} 2|V_t - M_t^V| > \sum_{t=1}^{T} |V_t - M_t^V| + \sum_{f=1}^{F} |V_f - M_f^V| \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$\sum_{t=1}^{T} |V_t - M_t^V| + \sum_{f=1}^{F} |V_f - M_f^V| = \sum_{j=1}^{N} |V_j - M_j^V|,$$

we can collect the terms on the right:

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} 2|V_t - M_t^V| > \sum_{j=1}^{N} |V_j - M_j^V| \\ 0 & \text{otherwise.} \end{cases}$$

After dividing both sides by the RHS term and dividing both sides by 2, we obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \frac{\sum_{t=1}^{T} |V_t - M_t^V|}{\sum_{j=1}^{N} |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

which is identical to

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} \frac{|V_t - M_t^V|}{\sum_{j=1}^{N} |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Since $V_t = 1$, we can multiply both sides by V_t and simplify the terms on the right to obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} \frac{|V_t - M_t^V| V_t}{\sum_{j=1}^{N} |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

As $V_f = 0$,

$$\sum_{f=1}^{F} \frac{|V_f - M_f^V| V_f}{\sum_{j=1}^{N} |V_f - M_f^V|} = 0,$$

and we can add this summation term to both sides of the previous equation and simplify the terms on the right to obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} \frac{|V_t - M_t^V| V_t}{\sum_{j=1}^{N} |V_j - M_j^V|} + \sum_{f=1}^{F} \frac{|V_f - M_f^V| V_f}{\sum_{j=1}^{N} |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Collecting the terms on the left, we obtain the weighted version of the SP algorithm, thus proving Lemma 1,

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} \frac{|V_i - M_i^V| V_i}{\sum_{j=1}^{N} |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Lemma 2: For MLRP to be satisfied, we need to show that for any set of signals that can be drawn with $s_i > s_j$ and $s_k > s_l$,

$$p(s_i|s_k)p(s_j|s_l) > p(s_j|s_k)p(s_i|s_l) \tag{6}$$

Note that a signal can only be drawn if it occurs with positive probability in its information service. Thus $p(s_i) > 0$, $p(s_j) > 0$, $p(s_k) > 0$, and $p(s_l) > 0$.

Assumption 2 implies that

$$p(s_a|s_b) = \frac{p(s_a, s_b)}{p(s_b)} = \frac{p(s_a|T)p(s_b|T)p(T) + p(s_a|F)p(s_b|F)p(F)}{p(s_b)}.$$

For $a \in \{i, j\}$ and $b \in \{k, l\}$. Rearranging Bayes Rule, it is the case that:

$$\frac{p(T)p(s_b|T)}{p(s_b)} = p(T|s_b) = s_b$$

and thus

$$p(s_a|s_b) = p(s_a|T)s_b + p(s_a|F)(1 - s_b) = Q^t_{Ta}s_b + Q^t_{Fa}(1 - s_b). \tag{7}$$

We first prove that MLRP holds for the case where $s_i > s_j > 0$ and $s_k > s_l > 0$. By the construction of the Q matrix

$$s_a = \frac{Q^t_{Ta}p(T)}{Q^t_{Ta}p(T) + Q^t_{Fa}(1 - p(T))}.$$

Under the assumption that $s_j > 0$ and $s_l > 0$, this can be rewritten as

$$Q^t_{Fa} = \frac{1 - s_a}{s_a}\frac{p(T)}{1 - p(T)}Q^t_{Ta} = \frac{1 - s_a}{s_a}\frac{s_\varnothing}{1 - s_\varnothing}Q^t_{Ta}$$

for $a \in \{i, j\}$. Substituting this into (7) implies that

$$p(s_a|s_b) = Q^t_{Ta}s_b\left[1 + \frac{1 - s_a}{s_a}\frac{1 - s_b}{s_b}\frac{s_\varnothing}{1 - s_\varnothing}\right]. \tag{8}$$

Let $r_{ab} = \frac{1-s_a}{s_a}\frac{1-s_b}{s_b}\frac{s_\varnothing}{1-s_\varnothing}$ for $a \in \{i, j\}$ and $b \in \{k, l\}$. Substituting (8) into (6), MLRP is satisfied if:

$$(1 + r_{ik})(1 + r_{jl}) > (1 + r_{il})(1 + r_{jk})$$

Expanding this equation, MLRP is satisfied if:

$$1 + r_{ik} + r_{jl} + r_{ik}r_{jl} > 1 + r_{il} + r_{jk} + r_{il}r_{jk}$$

Next, noting that $r_{ik}r_{jl} = r_{il}r_{jk}$, MLRP is satisfied if

$$r_{ik} + r_{jl} > r_{il} + r_{jk}.$$

Rearranging this equation, MLRP is satisfied if

$$\left[\frac{1 - s_i}{s_i} - \frac{1 - s_j}{s_j}\right]\left[\frac{1 - s_k}{s_k} - \frac{1 - s_l}{s_l}\right] > 0.$$

By the assumption that $s_i > s_j$ and $s_k > s_l$, both terms on the LHS are negative and thus this equation always holds.

We now check the cases for which (i) $s_j = 0$ but $s_l > 0$, (ii) $s_j > 0$ but $s_l = 0$, and (iii) both $s_j = 0$ and $s_l = 0$. When $s_j = 0$ but $s_l > 0$, $p(s_j|s_b) = Q^t_{Fj}(1 - s_b)$ and MLRP is satisfied if

$$[Q^t_{Ti}s_k(1 + r_{ik})][Q^t_{Fj}(1 - s_l)] > [Q^t_{Ti}s_l(1 + r_{il})][Q^t_{Fj}(1 - s_k)].$$

When $s_k = 1$, the RHS is zero and the LHS is strictly positive. Thus MLRP holds. When $s_k < 1$, the equation is equivalent to

$$\frac{s_k}{1 - s_k} > \frac{s_l}{1 - s_l},$$

which is satisfied due to the assumption that $s_k > s_l$.

When $s_l = 0$ and $s_j > 0$, MLRP holds if

$$\frac{Q^t_{Ti}s_k + Q^t_{Fi}(1 - s_k)}{Q^t_{Tj}s_k + Q^t_{Fj}(1 - s_k)} > \frac{Q^t_{Fi}}{Q^t_{Fj}}. \tag{9}$$

If $s_i = 1$, the RHS is equal to zero and the LHS is strictly positive. Thus MLRP holds. When $s_i < 1$, $s_j > 0$, and $s_l = 0$, $Q^t_{Ti} = \frac{s_i}{1-s_i}\frac{1-s_\varnothing}{s_\varnothing}Q^t_{Fi}$ and, after some algebra, the equation is equivalent to

$$\frac{s_i}{1 - s_i} > \frac{s_j}{1 - s_j},$$

which is always satisfied. Thus MLRP holds in this case.

Finally when $s_l = 0$ and $s_j = 0$, MLRP holds if

$$[Q^t_{Ti}s_k + Q^t_{Fi}(1 - s_k)]Q^t_{Fj} > [Q^t_{Fj}(1 - s_k)]Q^t_{Fi}.$$

Rearranging, MLRP holds if $Q^t_{Ti}s_kQ^t_{Fj} > 0$, which is always true.

Proof of Lemma 3: Assume that the event is true. The share of true votes from information service $t \in \{E, N\}$ (given that the state is true) is given by

$$\mathbb{E}V(Q^t|T) = \sum_{\{i|s_i \geq .5\}} \gamma(Q^t_{Ti}),$$

where $\gamma(Q^t_{oi}) = \frac{1}{2}Q^t_{oi}$ if $s_i = .5$ and $\gamma(Q^t_{oi}) = Q^t_{oi}$ otherwise. The meta-prediction of an individual in group t with signal s_k is

$$M^V(Q^t|s_k) = s_k\mathbb{E}V(Q^t|T) + (1 - s_k)\mathbb{E}V(Q^t|F).$$

The expected meta-prediction of forecasters in information service Q^t made by forecasters with information service Q^τ ($\tau = \{N, E\}$) when the state is o is given by

$$\mathbb{E}M^V(Q^t|Q^\tau, o) = \sum_k M^V(Q^t|s_k)Q^\tau_{ok}.$$

Aggregating up across novices and experts, the expected meta-prediction of votes from information service Q^t given state o is

$$\mathbb{E}M^V(Q^t|o) = \theta\mathbb{E}M^V(Q^t|Q^E, o) + (1 - \theta)\mathbb{E}M^V(Q^t|Q^N, o).$$

In the true state, the meta-prediction will underestimate (or be equal to) the true proportion of votes for the true state if for all $t \in \{E, N\}$,

$$\mathbb{E}V(Q^t|T) \geq \mathbb{E}M^V(Q^t|T). \tag{10}$$

We allow for equality here to account for the cases where (i) all individuals know the state is true or (ii) all individuals are uninformed with a prior of $s_\varnothing = .5$. In these special cases the expected votes and expected meta predictions will be equal.

Noting that $\mathbb{E}V(Q^t|T) = \theta\mathbb{E}V(Q^t|T) + (1-\theta)\mathbb{E}V(Q^t|T)$, equation (10) holds if for $t \in \{E, N\}$ and $\tau \in \{E, N\}$,

$$\mathbb{E}V(Q^t|T) \geq \mathbb{E}M^V(Q^t|Q^\tau, T)$$

Next, noting that (i)

$$\mathbb{E}M^V(Q^t|Q^\tau, T) = \sum_k M^V(Q^t|s_k)Q^\tau_{Tk},$$

(ii) $M^V(Q^t|s_k) = s_k\mathbb{E}V(Q^t|T) + (1-s_k)\mathbb{E}V(Q^t|F)$, and (iii) $\sum_k Q^\tau_{Tk} = 1$, equation (10) holds if

$$\sum_k (1-s_k)[\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F)]Q^\tau_{Tk} \geq 0.$$

This will be satisfied if $\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F) \geq 0$ for all t. This is equivalent to requiring that

$$\sum_{\{i|s_i \geq .5\}} [\gamma(Q^t_{Ti}) - \gamma(Q^t_{Fi})] \geq 0. \tag{11}$$

Noting that an information service is a stochastic matrix and that the rows add up to one, (11) is satisfied if

$$\sum_{\{i|s_i \leq .5\}} [\gamma(Q^t_{Fi}) - \gamma(Q^t_{Ti})] \geq 0. \tag{12}$$

Define the cumulative density function of $p(\hat{s}|s_b)$ as

$$G(\hat{s}|s_b) = \sum_{\{a|s_a \leq \hat{s}\}} p(s_a|s_b),$$

where $\hat{s} \in \{s_1, \ldots, s_m\} \cup \{s_\varnothing\}$. By lemma 1, MLRP holds. This implies that

$$p(\hat{s}|s_k)p(s_j|s_l) > p(s_j|s_k)p(\hat{s}|s_l)$$

for all $s_j < \hat{s}$. Noting that $p(\hat{s}|s_k)p(s_j|s_l) = p(s_j|s_k)p(\hat{s}|s_l)$ when $s_j = \hat{s}$,

$$p(\hat{s}|s_k)p(s_j|s_l) \geq p(s_j|s_k)p(\hat{s}|s_l)$$

for all $s_j \leq \hat{s}$. Summing both sides of this equation from s_0 to $\hat{s}$ with respect to s_j, MLRP implies

$$\frac{p(\hat{s}|s_k)}{p(\hat{s}|s_l)} \geq \frac{G(\hat{s}|s_k)}{G(\hat{s}|s_l)}$$

for all $\hat{s}$. MLRP also implies that

$$p(s_i|s_k)p(\hat{s}|s_l) \geq p(\hat{s}|s_k)p(s_i|s_l)$$

for all $s_i \geq \hat{s}$. Summing both sides of this equation over all $s_i > \hat{s}$, MLRP implies

$$\frac{1-G(\hat{s}|s_k)}{1-G(\hat{s}|s_l)} \geq \frac{p(\hat{s}|s_k)}{p(\hat{s}|s_l)}.$$

Combining these two equations we have

$$\frac{1-G(\hat{s}|s_k)}{1-G(\hat{s}|s_l)} \geq \frac{G(\hat{s}|s_k)}{G(\hat{s}|s_l)}$$

or

$$\frac{G(\hat{s}|s_l)}{1-G(\hat{s}|s_l)} \geq \frac{G(\hat{s}|s_k)}{1-G(\hat{s}|s_k)},$$

which implies $G(\hat{s}|s_l) \geq G(\hat{s}|s_k)$ for any $\hat{s}$ and for signals $s_l < s_k$.

When $s_l = 0$, $G(\hat{s}|0) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Fi}$. Further, when $s_k = 1$, $G(\hat{s}|1) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Ti}$. Thus MLRP implies

$$\sum_{\{i|s_i \leq \hat{s}\}} Q^t_{Ti} \leq \sum_{\{i|s_i \leq \hat{s}\}} Q^t_{Fi}$$

and thus equation (12) holds. The proof for the case where the event is false uses the same logic as the case where the state is true. In this case

$$\mathbb{E}M^V(Q^t|Q^\tau, F) = \sum_k M^V(Q^t|s_k) Q^\tau_{Fk},$$

and expanding $M^V(Q^t|s_k) = s_k \mathbb{E}V(Q^t|T) + (1-s_k)\mathbb{E}V(Q^t|F)$,

$$\mathbb{E}M^V(Q^t|Q^\tau, F) - \mathbb{E}V(Q^t|F) = \sum_k [s_k(\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F))] Q^\tau_{Fk}.$$

This is greater or equal to zero if $\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F) \geq 0$ for all Q^t. We have shown this to be true by MLRP above.

Proof of Lemma 4: Assume that the event is true. The expected probability prediction of forecasters from information service $t \in \{E, N\}$ in state $o \in \{T, F\}$ is given by

$$\mathbb{E}P(Q^t|o) = \sum_{s_i} s_i Q^t_{oi}.$$

The meta-prediction of an individual in group t with signal s_k is

$$M^P(Q^t|s_k) = s_k \mathbb{E}P(Q^t|T) + (1-s_k)\mathbb{E}P(Q^t|F).$$

The expected meta-prediction of forecasters in information service Q^t made by forecasters with information service Q^τ ($\tau \in \{N, E\}$) when the state is o is given by

$$\mathbb{E}M^P(Q^t|Q^\tau, o) = \sum_k M^P(Q^t|s_k) Q^\tau_{ok}.$$

Aggregating up across novices and experts, the expected probability meta-prediction from information service Q^t given state o is

$$\mathbb{E}M^P(Q^t|o) = \theta \mathbb{E}M^P(Q^t|Q^E, o) + (1-\theta)\mathbb{E}M^P(Q^t|Q^N, o).$$

In the true state, the probability meta-prediction will underestimate the true probability average if for all $t \in \{E, N\}$,

$$\mathbb{E}P(Q^t|T) \geq \mathbb{E}M^P(Q^t|T). \tag{13}$$

We again allow for equality here to account for cases where (i) all individuals know the state is true or (ii) all individuals receive $s_\varnothing$. In these special cases the probability meta-prediction will be equal to the average probability.

Noting that $\mathbb{E}P(Q^t|T) = \theta\mathbb{E}P(Q^t|T) + (1-\theta)\mathbb{E}P(Q^t|T)$, equation (13) holds if for $t \in \{E, N\}$ and $\tau \in \{E, N\}$,

$$\mathbb{E}P(Q^t|T) \geq \mathbb{E}M^P(Q^t|Q^\tau, T).$$

Next, recalling that (i)

$$\mathbb{E}M^P(Q^t|Q^\tau, o) = \sum_k M^P(Q^t|s_k)Q^\tau_{ok},$$

(ii) $M^P(Q^t|s_k) = s_k\mathbb{E}P(Q^t|T) + (1-s_k)\mathbb{E}P(Q^t|F)$, and (iii) $\sum_k Q^t_{Tk} = 1$, equation (13) holds if

$$\sum_k (1-s_k)[\mathbb{E}P(Q^t|T) - \mathbb{E}P(Q^t|F)] \geq 0.$$

This will be satisfied if $\mathbb{E}P(Q^t|T) - \mathbb{E}P(Q^t|F) \geq 0$ for all t.

Using the notation from Lemma 3, let $G(\hat{s}|0) = \sum_{\{i|s_i \leq \hat{s}\}} Q^t_{Fi}$ and $G(\hat{s}|1) = \sum_{\{i|s_i \leq \hat{s}\}} Q^t_{Ti}$ and recall that MLRP implies that for any $\hat{s}$

$$G(\hat{s}|0) \leq G(\hat{s}|1).$$

Thus $G(\hat{s}|1)$ First-order stochastic dominates $G(\hat{s}|0)$. An equivalent definition of stochastic dominance is that for any increasing function $u(\hat{s})$,

$$\sum_i u(s_i)Q_{Ti} \geq \sum_i u(s_i)Q_{Fi}$$

Using $u(\hat{s}) = \hat{s}$, this immediately implies that

$$\sum_i s_i[Q_{Ti} - Q_{Fi}] \geq 0,$$

which is equivalent to $\mathbb{E}P(Q^t|T) - \mathbb{E}P(Q^t|F)] \geq 0$. The proof for the case where the event is false uses the same logic as the case where the state is true. In this case we want to prove that $\mathbb{E}P(Q^t|F) \leq \mathbb{E}M^P(Q^t|Q^\tau, F)$. By definition,

$$\mathbb{E}M^P(Q^t|Q^\tau, F) = \sum_k M^P(Q^t|s_k)Q^\tau_{Fk}.$$

Expanding out $M^P(Q^t|s_k)Q^\tau_{Fk}$ and using the same steps as above, $\mathbb{E}P(Q^t|F) \leq \mathbb{E}M^P(Q^t|Q^\tau, F)$ if

$$\sum_k (1-s_k)[\mathbb{E}P(Q^t|F) - \mathbb{E}P(Q^t|T)] \leq 0.$$

We have shown by MLRP that $\mathbb{E}P(Q^t|F) \leq \mathbb{E}P(Q^t|T)$ and thus the condition holds for all k.

Propositions (1) – (6)

Proof of Proposition 1: In this proof, we show that in the SP algorithm, if (i) forecaster i is better-informed than forecaster j and (ii) the prior is unbiased, then the weight given to forecaster i will be strictly less than the weight given to forecaster j.

To begin, note that the when conditions (i) and (ii) hold above, either $\sigma_i < \sigma_j < 0.5$ or $\sigma_i > \sigma_j > 0.5$. Thus $V_i = V_j$. Without loss of generality, we concentrate on the case where $\sigma_i > \sigma_j > 0.5$ so that $V_i = V_j = 1$.

We are interested in the sign of $W^{SP}_i(\sigma_i) - W^{SP}_j(\sigma_j)$. If the sign is positive, then weights are increasing in signal and if it is negative, then weights will be decreasing in signal. Noting that the denominators of

$W_i^{SP}(\sigma_i)$ and $W_j^{SP}(\sigma_j)$ are positive and identical, the sign of $W_i^{SP}(\sigma_i) - W_j^{SP}(\sigma_j)$ will be the same as the sign of $|V(\sigma_i) - M^V(Q|\sigma_i)| - |V(\sigma_j) - M^V(Q|\sigma_j)|$.

As noted in the main text

$$M^V(Q|\sigma_k) = \sigma_k \mathbb{E}V(Q|T) + (1-\sigma_k)\mathbb{E}V(Q|F).$$

Thus, in the case where $\sigma_i > \sigma_j > 0.5$,

$$|V(\sigma_i) - M^V(Q|\sigma_i)| = 1 - \sigma_i \mathbb{E}V(Q|T) - (1-\sigma_i)\mathbb{E}V(Q|F)$$

and thus

$$\begin{aligned} |V(\sigma_i) - M^V(Q|\sigma_i)| - |V(\sigma_j) - M^V(Q|\sigma_j)| &= [\sigma_j - \sigma_i]\mathbb{E}V(Q|T) + [(1-\sigma_j) - (1-\sigma_i)]\mathbb{E}V(Q|F) \\ &= [\sigma_j - \sigma_i][\mathbb{E}V(Q|T) - \mathbb{E}V(Q|F)]. \end{aligned}$$

As shown in the proof of Lemma 3, $\mathbb{E}V(Q|T) > \mathbb{E}V(Q|F)$. Thus, since $\sigma_i > \sigma_j$, the sign of $|V(\sigma_i) - M^V(Q|\sigma_i)| - |V(\sigma_i) - M^V(Q|\sigma_i)|$ is negative. Thus, the weights given to forecaster i will be strictly less than the weight given to forecaster j.

Proof of Proposition 2: In this proof, we show that in the SC algorithm, if (i) forecaster i is better-informed than forecaster j, the weight given to i will be strictly more than the weight given to forecaster j.

Consider the case where $\sigma_i > \sigma_j > s_\varnothing$ We are interested in the sign of $W_i^{SC}(\sigma_i) - W_j^{SC}(\sigma_j)$. If the sign is positive, then weights are increasing in signal and if it is negative, then weights will be decreasing in signal. Noting that the denominators of $W_i^{SP}(\sigma_i)$ and $W_j^{SP}(\sigma_j)$ are positive and identical, the sign of $W_i^{SP}(\sigma_i) - W_j^{SP}(\sigma_j)$ will be the same as the sign of $|P(\sigma_i) - M^P(Q|\sigma_i)| - |P(\sigma_j) - M^P(Q|\sigma_j)|$.

As noted in the main text

$$M^P(Q|\sigma_k) = \sigma_k \mathbb{E}P(Q|T) + (1-\sigma_k)\mathbb{E}P(Q|F)$$

Thus, in the case where $\sigma_i > \sigma_j > 0.5$,

$$|P(\sigma_i) - M^P(Q|\sigma_i)| = \sigma_i - \sigma_i \mathbb{E}P(Q|T) - (1-\sigma_i)\mathbb{E}P(Q|F)$$

and thus

$$\begin{aligned} |P(\sigma_i) - M^P(Q|\sigma_i)| - |P(\sigma_j) - M^P(Q|\sigma_j)| &= [\sigma_i - \sigma_j] - [\sigma_i - \sigma_j]\mathbb{E}P(Q|T) + [\sigma_i - \sigma_j]\mathbb{E}P(Q|F) \\ &= [\sigma_i - \sigma_j][1 - (\mathbb{E}P(Q|T) - \mathbb{E}P(Q|F))]. \end{aligned}$$

As shown in the proof of Lemma 4, $0 \leq \mathbb{E}P(Q|F) < \mathbb{E}P(Q|T) \leq 1$. Thus, $[1 - (\mathbb{E}P(Q|T) - \mathbb{E}P(Q|F))]$ is positive. Since, $\sigma_i > \sigma_j$, the sign of $|P(\sigma_i) - M^P(Q|\sigma_i)| - |P(\sigma_i) - M^P(Q|\sigma_i)|$ is positive. As a consequence, the weights given to forecaster i will be strictly greater than the weight given to forecaster j.

Proof of Proposition 3: Define the absolute value of the expected contribution of a forecaster who receives signal s_i in the SP algorithm as:

$$|C^{SP}(Q|s_i)| = \begin{cases} |-M_i^V(s_i)| & \text{if } s_i < 0.5, \\ \frac{1}{2}|-M_i^V(s_i)| + \frac{1}{2}|1-M_i^V(s_i)| & \text{if } s_i = 0.5, \\ |1-M_i^V(s_i)| & \text{if } s_i > 0.5. \end{cases}$$

Summing up over forecasters, let $\frac{1}{N}\sum_i^N \mathbb{I}_{\{\sigma_i=s_i\}}$ be the proportion of forecasters that receives signal s_i. Then, Borel's law of large numbers implies that with probability one,

$$\frac{1}{N}\sum_i^N \mathbb{I}_{\{\sigma_i=s_j\}} \to Q_{oj} \text{ as } N \to \infty$$

for $o \in \{T, F\}$ and for all s_j. Since each forecaster receives an independent signal, this result implies that

$$\lim_{N\to\infty} \frac{1}{N}\sum_i^N |V_i - M_i^V| = \lim_{N\to\infty} \frac{1}{N}\sum_{s_j}\left(\sum_i^N |V_i - M_i^V|\mathbb{I}_{\{\sigma_i=s_j\}}\right) = \sum_{s_j} |C^{SP}(Q|s_j)|Q_{oj}$$

and

$$\lim_{N\to\infty} \frac{1}{N}\sum_i^N |V_i - M_i^V|V_i = \sum_{\{s_j|s_j\geq 0.5\}} |1-M^V(Q|s_j)|\gamma(Q_{oj})$$

Combining these two results,

$$\lim_{N\to\infty} \sum_i^N W_i^{SP}V_i = \lim_{N\to\infty} \frac{\frac{1}{N}\sum_i^N |V_i - M_i^V|V_i}{\frac{1}{N}\sum_i^N |V_i - M_i^V|} = \frac{\sum_{\{s_j|s_j\geq 0.5\}} |1-M^V(Q|s_j)|\gamma(Q_{oj})}{\sum_{s_j} |C^{SP}(Q|s_j)|Q_{oj}}. \tag{14}$$

Likewise, in the SC algorithm,

$$\lim_{N\to\infty} \frac{1}{N}\sum_i^N |P_i - M_i^P| = \sum_{s_j} |s_j - M^P(s_j)|Q_{oj}$$

and

$$\lim_{N\to\infty} \frac{1}{N}\sum_i^N |P_i - M_i^P|\mathbb{I}_{\{P_i>M_i^P\}} = \sum_{\{s_j|s_j\geq s_\varnothing\}} |s_j - M^P(s_j)|Q_{oj}.$$

Combining these results,

$$\lim_{N\to\infty} \sum_i^N W_i^{SC}\mathbb{I}_{\{P_i>M_i^P\}} = \lim_{N\to\infty} \frac{\frac{1}{N}\sum_i^N |P_i - M_i^P|\mathbb{I}_{\{P_i>M_i^P\}}}{\frac{1}{N}\sum_i^N |P_i - M_i^P|} = \frac{\sum_{\{s_j|s_j\geq s_\varnothing\}} |s_j - M^P(s_j)|Q_{oj}}{\sum_{s_j} |s_j - M^P(s_j)|Q_{oj}}. \tag{15}$$

Equations (14) and (15) imply that

$$\lim_{N\to\infty} \sum_i^N W_i^{SC}\mathbb{I}_{\{P_i>M_i^P\}} \geq \lim_{N\to\infty} \sum_i^N W_i^{SP}V_i$$

if and only if

$$\frac{\sum_{\{s_j|s_j\geq s_\varnothing\}} |s_j - M^P(s_j)|Q_{oj}}{\sum_{s_j} |s_j - M^P(s_j)|Q_{oj}} \geq \frac{\sum_{\{s_j|s_j\geq 0.5\}} |1-M^V(Q|s_j)|\gamma(Q_{oj})}{\sum_{s_j} |C^{SP}(Q|s_j)|Q_{oj}} \tag{16}$$

for $o \in \{T, F\}$. We will prove this relationship holds in the true state when the problem is unbiased and $s_\varnothing = 0.5$. The case for the false state is identical and is omitted.

Starting with the left hand side of equation (16),

$$\sum_{\{s_j|s_j\geq 0.5\}} |s_j - M^P(s_j)|Q_{Tj} = \sum_{\{s_j|s_j\geq 0.5\}} |s_j - s_j\mathbb{E}(P|T) - (1-s_j)\mathbb{E}(P|F)|Q_{Tj}$$

By the law of total probability $\mathbb{E}(P) = 0.5\mathbb{E}(P|T) + 0.5\mathbb{E}(P|F)$. Noting that $\mathbb{E}(P) = 0.5$ in the unbiased case, this implies $E(P|F) = 1 - E(P|T)$. Thus:

$$\sum_{\{s_j|s_j \geq 0.5\}} |s_j - M^P(s_j)|Q_{Tj} = \sum_{\{s_j|s_j \geq 0.5\}} |2s_j - 1|[1 - \mathbb{E}(P|T)]Q_{Tj}.$$

Likewise,

$$\sum_{s_j} |s_j - M^P(s_j)|Q_{Tj} = \sum_{s_j} |2s_j - 1|[1 - \mathbb{E}(P|T)]Q_{Tj}.$$

Thus, the left hand side of equation (16) in the true state with an unbiased prior is equal to:

$$\frac{\sum_{\{s_j|s_j \geq 0.5\}} |2s_j - 1|Q_{Tj}}{\sum_{s_j} |2s_j - 1|Q_{Tj}}. \tag{17}$$

We can also simplify the right hand side of (16). By the law of total probability, $\mathbb{E}(V) = .5\mathbb{E}(V|T) + .5\mathbb{E}(V|F)$, and thus $\mathbb{E}(V|F) = 2\mathbb{E}(V) - \mathbb{E}(V|T)$. Thus, for $s_j \geq 0.5$:

$$\begin{aligned} |1 - M^V(Q|s_j)| &= 1 - s_j\mathbb{E}(V|T) - (1 - s_j)[2\mathbb{E}(V) - \mathbb{E}(V|T)] \\ &= 1 - \mathbb{E}(V) - (2s_j - 1)[\mathbb{E}(V|T) - \mathbb{E}(V)] \\ &= 1 - \mathbb{E}(V) - |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]. \end{aligned}$$

Likewise, for $s_j < 0.5$:

$$\begin{aligned} |-M^V(Q|s_j)| &= |-s_j\mathbb{E}(V|T) - (1 - s_j)[2\mathbb{E}(V) - \mathbb{E}(V|T)]| \\ &= |-\mathbb{E}(V) + (1 - 2s_j)[\mathbb{E}(V|T) - \mathbb{E}(V)]| \\ &= \mathbb{E}(V) - |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]. \end{aligned}$$

Finally, if $s_j = 0.5$, then

$$|C^{SP}(Q|s_j)| = \frac{1}{2}[1 - \mathbb{E}(V)] + \frac{1}{2}\mathbb{E}(V).$$

Noting that when $s_j = 0.5$, $\frac{1}{2}[1 - \mathbb{E}(V)]Q_{Ti} = [1 - \mathbb{E}(V)]\gamma(Q_{Ti})$ and $\frac{1}{2}[\mathbb{E}(V)]Q_{Ti} = [\mathbb{E}(V)]\gamma(Q_{Ti})$, the right hand side of (16) can be rewritten as

$$\frac{\sum_{\{s_j|s_j \geq .5\}}[1 - E(V)]\gamma(Q_{Tj}) - \sum_{\{s_j|s_j \geq .5\}} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]\gamma(Q_{Tj})}{\sum_{\{s_j|s_j \leq .5\}}[E(V)]\gamma(Q_{Tj}) + \sum_{\{s_j|s_j \geq .5\}}[1 - E(V)]\gamma(Q_{Tj}) + \sum_{s_j \neq 0.5} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]Q_{Tj}}.$$

This is equivalent to

$$\frac{\sum_{\{s_j|s_j \geq .5\}}[1 - E(V)]\gamma(Q_{Tj}) - \sum_{\{s_j|s_j \geq .5\}} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]Q_{Tj}}{\sum_{\{s_j|s_j \leq .5\}}[E(V)]\gamma(Q_{Tj}) + \sum_{\{s_j|s_j \geq .5\}}[1 - E(V)]\gamma(Q_{Tj}) + \sum_{s_j} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]Q_{Tj}} \tag{18}$$

since $|2s_j - 1| = 0$ for $s_j = 0.5$ and $\gamma(Q_{Tj}) = Q_{Tj}$ for $s_j \neq 0.5$.

Noting that $\sum_{\{s_j|s_j \geq .5\}}[1 - E(V)]\gamma(Q_{Ti}) = [1 - \mathbb{E}(V)]\mathbb{E}(V|T)$, and $\sum_{\{s_j|s_j \leq .5\}}[E(V)]\gamma(Q_{Ti}) = \mathbb{E}(V)[1 - \mathbb{E}(V|T)]$, equation (18) can be rewritten as:

$$\frac{\frac{\mathbb{E}(V|T)[1-\mathbb{E}(V)]}{\mathbb{E}(V|T)-\mathbb{E}(V)] } - \sum_{\{s_j|s_j \geq .5\}} |2s_j - 1|Q_{Tj}}{\frac{[1-\mathbb{E}(V|T)]\mathbb{E}(V)+\mathbb{E}(V|T)[1-\mathbb{E}(V)]}{\mathbb{E}(V|T)-\mathbb{E}(V)} - \sum_{s_j} |2s_j - 1|Q_{Tj}}. \tag{19}$$

Cross multiplication shows that for any values of x, y, a, and b with $x > b > 0$ and $y > a > 0$,

$$\frac{a}{b} \geq \frac{x - a}{y - b}$$

if and only if

$$\frac{a}{b} \geq \frac{x}{y}$$

Thus, to show that equation (17) is greater than equation (19), it is sufficient to show that

$$\frac{\sum_{\{s_j|s_j\geq 0.5\}} |2s_j-1|Q_{Tj}}{\sum_{s_j} |2s_j-1|Q_{Tj}} \geq \frac{\mathbb{E}(V|T)[1-\mathbb{E}(V)]}{\mathbb{E}(V|T)[1-\mathbb{E}(V)]+[1-\mathbb{E}(V|T)]\mathbb{E}(V)}. \tag{20}$$

Next, note that

$$\begin{aligned}\sum_{\{s_j|s_j\geq 0.5\}} |2s_j-1|Q_{Tj} &= \sum_{\{s_j|s_j\geq 0.5\}} (2s_j-1)Q_{Tj} \\ &= 2\mathbb{E}(V|T)[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5]\end{aligned}$$

and

$$\begin{aligned}\sum_{s_j} |2s_j-1|Q_{Tj} &= \sum_{\{s_j|s_j\geq 0.5\}} (2s_j-1)Q_{Tj} + \sum_{\{s_j|s_j\leq 0.5\}} (1-2s_j)Q_{Tj} \\ &= 2\mathbb{E}(V|T)[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5] + 2[1-\mathbb{E}(V|T)][0.5-\mathbb{E}(s_j|s_j\leq 0.5,T)]\end{aligned}$$

Thus, we can rewrite the condition in (20) as

$$\frac{\mathbb{E}(V|T)[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5]}{\mathbb{E}(V|T)[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5]+[1-\mathbb{E}(V|T)][0.5-\mathbb{E}(s_j|s_j\leq 0.5,T)]} \geq \frac{\mathbb{E}(V|T)[1-\mathbb{E}(V)]}{\mathbb{E}(V|T)[1-\mathbb{E}(V)]+[1-\mathbb{E}(V|T)\mathbb{E}(V)}$$

or, equivalently

$$\frac{\mathbb{E}(V|T)\frac{[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5]}{[0.5-\mathbb{E}(s_j|s_j\leq 0.5,T)]}}{\mathbb{E}(V|T)\frac{[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5]}{[0.5-\mathbb{E}(s_j|s_j\leq 0.5,T)]}+[1-\mathbb{E}(V|T)]} \geq \frac{\mathbb{E}(V|T)\frac{[1-\mathbb{E}(V)]}{E(V)}}{\mathbb{E}(V|T)\frac{[1-\mathbb{E}(V)]}{E(V)}+[1-\mathbb{E}(V|T)]}.$$

Cross multiplication shows that for any x, y, and α with $x\geq 0$, $y\geq 0$, and $\alpha\in(0,1)$,

$$\frac{\alpha x}{\alpha x+(1-\alpha)} \geq \frac{\alpha y}{\alpha y+(1-\alpha)}$$

if and only if $x\geq y$. Thus, equation (20) is satisfied if

$$\frac{[\mathbb{E}(s_j|s_j\geq 0.5,T)-.5]}{0.5-\mathbb{E}(s_j|s_j\leq 0.5,T)} \geq \frac{[1-\mathbb{E}(V)]}{\mathbb{E}(V)}. \tag{21}$$

As a final step, note that

$$\mathbb{E}(s_j)=\mathbb{E}(s_j|s_j\geq .5)\mathbb{E}(V|T)+\mathbb{E}(s_j|s_j\leq .5)[1-\mathbb{E}(V|T)].$$

Since the decision problem is unbiased, $\mathbb{E}(s_j)=0.5$, and thus

$$[\mathbb{E}(s_j|s_j\geq .5)-.5]\mathbb{E}(V|T)=[.5-\mathbb{E}(s_j|s_j\leq .5)][1-\mathbb{E}(V|T)].$$

Rearranging this equation, it is the case that

$$\frac{[\mathbb{E}(s_j|s_j\geq 0.5)-.5]}{0.5-\mathbb{E}(s_j|s_j\leq 0.5)} = \frac{[1-\mathbb{E}(V)]}{\mathbb{E}(V)}$$

Finally, since $Q_{Tj}>Q_{Fj}$ for all $s_j>0.5$,

$$\mathbb{E}(s_j|s_j\geq 0.5,T)>\mathbb{E}(s_j|s_j\geq 0.5).$$

Likewise, $Q_{Tj} < Q_{Fj}$ for all $s_j < 0.5$. Thus

$$\mathbb{E}(s_j | s_j \leq 0.5, T) < \mathbb{E}(s_j | s_j \leq 0.5).$$

Thus, it is the case that

$$\frac{[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5]}{0.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)} \geq \frac{[\mathbb{E}(s_j | s_j \geq 0.5) - .5]}{0.5 - \mathbb{E}(s_j | s_j \leq 0.5)} = \frac{[1 - \mathbb{E}(V)]}{\mathbb{E}(V)}.$$

Thus equation (21) is satisfied. This implies that equation (20) is also satisfied and that equation (17) is greater than equation (19).

Proof of Proposition 4: A forecaster with signal s_k will make a probabilistic forecast of s_k. Thus, given an outcome state o, the expected prediction from information service Q^t is given by

$$P(Q^t | o) = \sum_{\{i | s_i \geq 0\}} Q^t_{oi} s_i.$$

Aggregating over both information services, the expected prediction of the population in state o is given by

$$\mathbb{E}P(\theta | o) := \theta \mathbb{E}P(Q^E | o) + (1 - \theta)\mathbb{E}P(Q^N | o)$$

In the absence of any information service, the probabilistic forecast of each individual would be $s_\varnothing$. By the law of total expectations, the posteriors are a mean-preserving spread of the prior, and thus we have

$$s_\varnothing = s_\varnothing \mathbb{E}P(Q^\tau | T) + (1 - s_\varnothing)\mathbb{E}P(Q^\tau | F).$$

for $\tau \in \{E, N\}$. This also implies that

$$s_\varnothing = s_\varnothing \mathbb{E}P(\theta | T) + (1 - s_\varnothing)\mathbb{E}P(\theta | F)$$

and that

$$\mathbb{E}P(\theta | F) = \frac{s_\varnothing}{1 - s_\varnothing}[1 - \mathbb{E}P(\theta | T)]. \tag{22}$$

A forecaster with signal s_k's meta-prediction about the others is equal to

$$M^P(\theta | s_k) = s_k \mathbb{E}P(\theta | T) + (1 - s_k)\mathbb{E}P(\theta | F).$$

Substituting in for $\mathbb{E}P(\theta | F)$ using (22), the meta-prediction of an individual with signal s_k can be expressed as

$$M^P(\theta | s_k) = s_k \mathbb{E}P(\theta | T) + (1 - s_k)\frac{s_\varnothing}{1 - s_\varnothing}[1 - \mathbb{E}P(\theta | T)].$$

The total contribution of an individual is based on the difference between the individual's prediction and meta-prediction. For an individual with signal s_k,

$$s_k - M^P(\theta | s_k) = s_k - s_k \mathbb{E}P(\theta | T) - (1 - s_k)\frac{s_\varnothing}{1 - s_\varnothing}[1 - \mathbb{E}P(\theta | T)]$$

or, equivalently:

$$s_k - M^P(\theta | s_k) = \frac{s_k - s_\varnothing}{1 - s_\varnothing}[1 - \mathbb{E}P(\theta | T)]. \tag{23}$$

Note first that the difference between an individual's signal and his or her meta-prediction is zero at $s_\varnothing$ and is linearly increasing in s_k. This feature implies that the weight of an individual with signal s_k, proportional to $|s_k - M^P(\theta|s_k)|$, is directly related to the informativeness of the posterior that an individual holds relative to the prior. Thus, individuals with more informative posteriors (an ex-post notion of expertise) will be weighted proportionally more than individuals with less informative posteriors.

We now show that $\mathbb{E}[P(Q^E|T)] - \mathbb{E}[M^P(\theta|Q^E,T)] \geq \mathbb{E}[P(Q^N|T)] - \mathbb{E}[M^P(\theta|Q^N,T)]$ First note that because $\sum_i Q^E_{Ti} = 1$

$$\begin{aligned}\mathbb{E}[P(Q^E|T)] - \mathbb{E}[M^P(\theta|Q^E,T)] &= \sum_i \left[\frac{s_i - s_\varnothing}{1 - s_\varnothing}[1 - \mathbb{E}P(\theta|T)]\right] Q^E_{Ti} \\ &= \frac{[1 - \mathbb{E}P(\theta|T)]}{1 - s_\varnothing}[(\sum_i s_i Q^E_{Ti}) - s_\varnothing].\end{aligned}$$

Thus, $\mathbb{E}[P(Q^E|T)] - \mathbb{E}[M^P(\theta|Q^E,T)] - \mathbb{E}[P(Q^N|T)] - \mathbb{E}[M^P(\theta|Q^N,T)]$ is equal to

$$\frac{[1 - \mathbb{E}P(\theta|T)]}{1 - s_\varnothing}[(\sum_i s_i Q^E_{Ti}) - (\sum_i s_i Q^N_{Ti})]$$

The sign of this equation will be positive if

$$\sum_i s_i Q^E_{Ti} \geq \sum_i s_i Q^N_{Ti},$$

or, equivalently if $\mathbb{E}[s_i|Q^E,T] \geq \mathbb{E}[s_i|Q^N,T]$.

We now show that $\mathbb{E}[s_i|Q^E,T] \geq \mathbb{E}[s_i|Q^N,T]$. To do so, we will use Blackwell's Theorem (Blackwell 1951):

Blackwell's Theorem *For information service Q^E to be more informative than Q^N it is necessary and sufficient that the value of information in service Q^E is greater than the value of information in service Q^N for all sets of terminal actions, all utility functions, and all prior beliefs.*

By Assumption 1, Q^E is more informative than Q^N. Let the action set $V \in \{T, F\}$ correspond to voting on whether an answer is true or false, and consider a utility function $U(V, o)$ that maps actions and states of the world into outcomes. Let $U(T,T) = 1$, $U(F,F) = 0$, $U(F,T) = 0$, and $U(T,F) = 0$. Given a signal s_i, expected utility is maximized by choosing $a = T$ in all states. The expected utility of this strategy given signal s_i is

$$\mathbb{E}[U(Q^t|s_i)] = U(T,T)s_i = s_i$$

By Blackwell's theorem, the expected utility of information service Q^E is higher than the expected value of information service Q^N for any utility function and any prior belief. Using an initial prior of $P(T) = 1$,

$$\mathbb{E}[U(Q^t)] = \sum_i \mathbb{E}[U(Q^t|s_i)]Q^t_{Ti} = \sum_i s_i Q^t_{Ti}$$

Thus, $\mathbb{E}[U(Q^E)] > \mathbb{E}[U(Q^N)]$ immediately implies

$$\sum_i s_i Q^E_{Ti} > \sum_i s_i Q^N_{Ti},$$

which implies that the sign of $\sum_i s_i Q^E_{Ti} - \sum_i s_i Q^N_{Ti}$ is positive.

The proof for the False state has an identical structure to the proof used for the True state. $\mathbb{E}[M^P(\theta|Q^E,F)]-\mathbb{E}[P(Q^E|F)]-\mathbb{E}[M^P(\theta|Q^N,T)]-\mathbb{E}[P(Q^N|T)]$ is equal to

$$\frac{[1-\mathbb{E}P(\theta|T)]}{1-s_\varnothing}[(\sum_i s_i Q^N_{Fi})-(\sum_i s_i Q^E_{Fi})]$$

The sign of this equation will be positive if

$$\sum_i -s_i Q^E_{Fi} > \sum_i -s_i Q^N_{Fi}.$$

or, equivalently,

$$\sum_i (1-s_i) Q^E_{Fi} > \sum_i (1-s_i) Q^N_{Fi}.$$

The left hand side of this last equation is $\mathbb{E}[1-s_i|Q^E,F]$ while the right hand side is $\mathbb{E}[1-s_i|Q^N,T]$. Using Blackwell's theorem with $U(F,F)=1$, $U(T,F)=U(F,T)=U(T,T)=0$ and $P(T)=0$ immediately shows that this condition holds.

Proof of Proposition 5: Assume that the event is true. The expected confidence prediction of forecasters from information service $t\in\{E,N\}$ in state $o\in\{T,F\}$ is given by

$$\mathbb{E}C(Q^t|o)=\sum_i\left(\sum_k c(s_k)R_{ik}\right)Q^t_{oi}.$$

By assumption, all forecasters' probability meta-predictions are fully adaptive. Thus, the (confidence adjusted) meta-prediction of an individual in group t with signal s_k is

$$M^C(Q^t|c(s_k))=c(s_k)\mathbb{E}C(Q^t|T)+(1-c(s_k))\mathbb{E}C(Q^t|F).$$

The expected meta-prediction of forecasters in information service Q^t made by forecasters with information service Q^τ ($\tau\in\{N,E\}$) when the state is o is given by

$$\mathbb{E}M^C(Q^t|Q^\tau,o)=\sum_i\left(\sum_k M^C(Q^t|c(s_k))R_{ik}\right)Q^\tau_{oi}.$$

Aggregating up across novices and experts, the expected probability meta-prediction from information service Q^t given state o is

$$\mathbb{E}M^C(Q^t|o)=\theta\mathbb{E}M^C(Q^t|Q^E,o)+(1-\theta)\mathbb{E}M^C(Q^t|Q^N,o).$$

In the true state, the probability meta-prediction will underestimate the true probability average if for all $t\in\{E,N\}$,

$$\mathbb{E}C(Q^t|T)\geq\mathbb{E}M^C(Q^t|T). \tag{24}$$

We allow for equality here to account for cases where (i) all individuals know the state is true or (ii) all individuals receive $s_\varnothing$. In these special cases the probability meta-prediction will be equal to the average probability.

Noting that $\mathbb{E}C(Q^t|T)=\theta\mathbb{E}C(Q^t|T)+(1-\theta)\mathbb{E}C(Q^t|T)$, equation (24) holds if for $t\in\{E,N\}$ and $\tau\in\{E,N\}$,

$$\mathbb{E}C(Q^t|T)\geq\mathbb{E}M^C(Q^t|Q^\tau,T).$$

Next, recalling that (i)

$$\mathbb{E}M^C(Q^t|Q^\tau,o)=\sum_i\left(\sum_k M^C(Q^t|c(s_k))R_{ik}\right)Q^\tau_{oi},$$

(ii) $M^C(Q^t|c(s_k))=c(s_k)\mathbb{E}C(Q^t|T)+(1-c(s_k))\mathbb{E}C(Q^t|F)$, and (iii) $\sum_i(\sum_k R_{ik})Q^\tau_{Ti}=1$, equation (24) holds if

$$\sum_i\left(\sum_k(1-c(s_k))[\mathbb{E}C(Q^t|T)-\mathbb{E}C(Q^t|F)]R_{ik}\right)Q^\tau_{Ti}\geq 0.$$

This will be satisfied if $\mathbb{E}C(Q^t|T)-\mathbb{E}C(Q^t|F)\geq 0$ for all t.

Noting that by part (iv) of the definition of systematically miscalibrated, $\sum_k c(s_k)R_{ik}=c(s_i)$, and thus

$$\mathbb{E}C(Q^t|o)=\sum_i c(s_i)Q^t_{oi}.$$

Thus, we need to show that $\sum_i c(s_i)Q^t_{Ti}>\sum_i c(s_i)Q^t_{Fi}$. Using the notation from Lemma 3, let $G(\hat{s}|0)=\sum_{\{i|s_i\leq\hat{s}\}}Q^t_{Fi}$ and $G(\hat{s}|1)=\sum_{\{i|s_i\leq\hat{s}\}}Q^t_{Ti}$ and recall that MLRP implies that for any $\hat{s}$

$$G(\hat{s}|0)\leq G(\hat{s}|1).$$

Thus $G(\hat{s}|1)$ first-order stochastic dominates $G(\hat{s}|0)$. An equivalent definition of stochastic dominance is that for any increasing function $u(\hat{s})$,

$$\sum_i u(s_i)Q_{Ti}\geq\sum_i u(s_i)Q_{Fi}$$

Using $u(\hat{s})=c(\hat{s})$, this immediately implies that

$$\sum_i c(s_i)[Q_{Ti}-Q_{Fi}]\geq 0,$$

which is equivalent to $\mathbb{E}C(Q^t|T)-\mathbb{E}C(Q^t|F)\geq 0$.

The proof for the case where the event is false uses the same logic as the case where the state is true. In this case we want to prove that $\mathbb{E}C(Q^t|F)\leq\mathbb{E}M^C(Q^t|Q^\tau,F)$. By definition,

$$\mathbb{E}M^C(Q^t|Q^\tau,F)=\sum_i\left(\sum_k M^C(Q^t|c(s_k))R_{ik}\right)Q^\tau_{Fi}.$$

Expanding out $M^C(Q^t|s_k)Q^\tau_{Fk}$ and using the same steps as above, $\mathbb{E}C(Q^t|F)\leq\mathbb{E}M^C(Q^t|Q^\tau,F)$ if

$$\sum_i\left(\sum_k(1-c(s_k))[\mathbb{E}C(Q^t|F)-\mathbb{E}C(Q^t|T)]R_{ik}\right)Q_{Fi}\leq 0.$$

We have shown by MLRP that $\mathbb{E}C(Q^t|F)\leq\mathbb{E}C(Q^t|T)$ and thus the condition holds for all t.

Proof of Proposition 6: Let $\rho^\tau_k:=Q^\tau_{Fk}+Q^\tau_{Tk}$. Then, by the assumption that $s_\varnothing=.5$, Bayes Rule implies

$$Q^\tau_{Tk}=s_k\rho^\tau_k$$

and, by definition,

$$\mathbb{E}V(Q^\tau|T)=\frac{1}{2}Q^\tau_{T\varnothing}+\sum_{\{k|s_k>.5\}}Q^\tau_{Tk}=\frac{1}{2}s_\varnothing\rho^\tau_\varnothing+\sum_{\{k|s_k>.5\}}s_k\rho^\tau_k.$$

Recall that $M^V(\theta|s_k)$ is defined as the probabilistic meta-prediction of a forecaster with signal s_k. Note that $M^V(\theta|s_k)$ is a weighted average of $M^V(Q^E|s_k)$ and $M^V(Q^N|s_k)$:

$$M^V(\theta|s_k) := \theta M^V(Q^E|s_k) + (1-\theta)M^V(Q^N|s_k).$$

Then, by definition,

$$\mathbb{E}M^V(\theta|Q^\tau,T) = \sum_k M^V(\theta|s_k)Q^\tau_{Tk} = \sum_{\{k|s_k<.5\}} M^V(\theta|s_k)s_k\rho^\tau_k + \frac{1}{2}s_\varnothing\rho^\tau_\varnothing + \sum_{\{k|s_k>.5\}} M^V(\theta|s_k)s_k\rho^\tau_k.$$

By the symmetry assumption, for all $k \le .5m$, (i) $s_k = 1 - s_{m+2-k}$, (ii) $M^V(\theta|s_k) = M^V(\theta|1-s_{m+2-k})$, and (iii) $\rho^\tau_k = \rho^\tau_{m+2-k}$. This implies

$$\begin{aligned}\sum_{\{k|s_k<.5\}} M^V(\theta|s_k)s_k\rho^\tau_k &= \sum_{\{k|s_{m+2-k}>.5\}} M^V(\theta|s_k)s_k\rho^\tau_k \\ &= \sum_{\{k|s_{m+2-k}>.5\}} M^V(\theta|1-s_{m+2-k})(1-s_{m+2-k})\rho^\tau_{m+2-k} \\ &= \sum_{\{l|s_l>.5\}} M^V(\theta|1-s_l)(1-s_l)\rho^\tau_l,\end{aligned}$$

where $l = m+2-k$. Noting that $l \in \{.5m+2,\dots,m+1\}$ and shares the same indexes as the original set $\{k|s_k>.5\}$, we can combine terms and rewrite

$$\mathbb{E}M^V(\theta|Q^\tau,T) = \frac{1}{2}s_k\rho^\tau_\varnothing + \sum_{\{k|s_k>.5\}} [M^V(\theta|s_k)s_k + M^V(\theta|1-s_k)(1-s_k)]\rho^\tau_k.$$

Using this representation of the meta prediction, the expected total contribution of an individual in group τ is:

$$\mathbb{E}V(Q^\tau|T) - \mathbb{E}M^V(\theta|Q^\tau,T) = \sum_{\{k|s_k>.5\}} [s_k - s_kM^V(\theta|s_k) - (1-s_k)M^V(\theta|1-s_k)]\rho^\tau_k. \tag{25}$$

By the definition of $M^V(\theta|s_k)$:

$$\begin{aligned}s_kM^V(\theta|s_k) + (1-s_k)M^V(\theta|1-s_k) &= \theta[s_kM^V(Q^E|s_k) + (1-s_k)M^V(Q^E|1-s_k)] \\ &\quad +(1-\theta)[s_kM^V(Q^N|s_k) + (1-s_k)M^V(Q^N|1-s_k)].\end{aligned} \tag{26}$$

Symmetry implies that $\mathbb{E}V(Q^\tau|F) = 1 - \mathbb{E}V(Q^\tau|T)$. Thus

$$\begin{aligned}M^V(Q^\tau|s_k) &= s_k\mathbb{E}V(Q^\tau|T) + (1-s_k)\mathbb{E}V(Q^\tau|F) \\ &= s_k\mathbb{E}V(Q^\tau|T) + (1-s_k)(1-\mathbb{E}V(Q^\tau|T)).\end{aligned}$$

This implies that for $s_k \ge .5$:

$$\begin{aligned}s_kM^V(Q^\tau|s_k) &= s_k[s_k\mathbb{E}V(Q^\tau|T) + (1-s_k)(1-\mathbb{E}V(Q^\tau|T))] \\ &= s_k(1-s_k) + (s_k^2 - s_k(1-s_k))\mathbb{E}V(Q^\tau|T)\end{aligned}$$

and

$$\begin{aligned}(1-s_k)M^V(Q^\tau|1-s_k) &= (1-s_k)[(1-s_k)\mathbb{E}V(Q^\tau|T) + s_k(1-\mathbb{E}V(Q^E|T)] \\ &= s_k(1-s_k) + ((1-s_k)^2 - s_k(1-s_k))\mathbb{E}V(Q^\tau|T)\end{aligned}$$

Substitution these two simplifications into (26) implies:

$$s_k M^V(\theta|s_k) + (1-s_k)M^V(\theta|1-s_k) = 2s_k(1-s_k) + (2s_k-1)^2[\theta \mathbb{E}V(Q^E|T) + (1-\theta)\mathbb{E}V(Q^N|T)] \tag{27}$$

Let $\mathbb{E}V(\theta|T) := \theta \mathbb{E}V(Q^E|T) + (1-\theta)\mathbb{E}V(Q^N|T)$ be the expected vote in the true state and note that this quantity is a constant. Then, using the simplification in (27), equation (25) implies

$$\begin{aligned}
\mathbb{E}V(Q^\tau|T) - \mathbb{E}M^V(\theta|Q^\tau,T) &= \sum_{\{k|s_k>.5\}} [s_k - 2s_k(1-s_k) - (2s_k-1)^2 \mathbb{E}V(\theta|T)]\rho_k^\tau \\
&= \sum_{\{k|s_k>.5\}} [(2s_k-1)(s_k(1-\mathbb{E}V(\theta|T)) + (1-s_k)\mathbb{E}V(\theta|T))]\rho_k^\tau \\
&= \left[\sum_{\{k|s_k>.5\}} \phi(s_k)\rho_k^\tau\right] + \phi(s_\varnothing)Q_{T\varnothing}^\tau,
\end{aligned}$$

where $\phi(s_k) = (2s_k-1)(s_k(1-\mathbb{E}V(\theta|T)) + (1-s_k)\mathbb{E}V(\theta|T))$ and $\phi(s_\varnothing) = 0$. Note that if a symmetric information service has only two posteriors that occur with positive probability, s_k and $(1-s_k)$, $\phi(s_k)$ is the expected difference between an individual's expected vote and their meta-prediction in the true state. This implies that when an information service is symmetric, the total contribution of a forecaster with information service Q^τ is the weighted average of simpler symmetric information services that contain only two posteriors.

To show that the expected total contribution of an expert is greater or equal to the expected total contribution of a novice, we need to show that

$$(\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E,T)) - (\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N,T)) \geq 0$$

or, equivalently,

$$\left[\sum_{\{k|s_k>.5\}} \phi(s_k)\rho_k^E\right] - \left[\sum_{\{l|s_l>.5\}} \phi(s_l)\rho_l^N\right] + \phi(s_\varnothing)[Q_{T\varnothing}^E - Q_{T\varnothing}^N] \geq 0$$

We do this in two steps. First, we construct a set of non-negative weights $w_{k,l}$ where (i) $w_{k,l} = 0$ in cases where $l > k$ and (ii)

$$\begin{aligned}
&\left[\sum_{\{k|s_k>.5\}} \phi(s_k)\rho_k^E\right] - \left[\sum_{\{l|s_l>.5\}} \phi(s_l)\rho_l^N\right] \\
&\qquad + \quad \phi(s_\varnothing)[Q_{T\varnothing}^E - Q_{T\varnothing}^N] = \sum_{\{k|s_k\geq.5\}} \sum_{\{l|s_l\geq.5\}} [\phi(s_k) - \phi(s_l)]w_{k,l}.
\end{aligned} \tag{28}$$

We then show that $\phi(s_k) - \phi(s_l) \geq 0$ for all $k > l$. This guarantees that each element in the RHS of (28) is positive or zero.

The assumption of strict garbling implies that

$$\sum_{\{i|s_i\leq\hat{s}\}} [Q_{Ti}^E + Q_{T(m+2-i)}^E] \geq \sum_{\{i|s_i\leq\hat{s}\}} [Q_{Ti}^N + Q_{T(m+2-i)}^N]$$

for all $\hat{s} \leq s_\varnothing$. Noting that $Q_{T(m+2-i)}^t = Q_{Fi}^t$, strict garbling implies

$$\sum_i \rho_i^E \geq \sum_i \rho_i^N$$

for all $i \in \{1, \ldots, \frac{m}{2}+1\}$. We use this relationship to construct a matrix of weights $W = [w_{ij}]_{(m+1)\times(m+1)}$ where (28) is satisfied.

We begin by constructing a submatrix consisting of the first $(\frac{m}{2}+1)\times(\frac{m}{2}+1)$ elements of W. Let

$$V^E = [\rho_1^E, \rho_2^E, \ldots, \rho_{\frac{m}{2}}^E, \frac{1}{2}\rho_\varnothing^E]$$

be a $\frac{m}{2}+1$ element vector. Note that $\frac{1}{2}\rho_\varnothing^E = Q_{T\varnothing}$ and thus, by construction, the elements of the vector sum to 1. Likewise, define

$$V^{N,1} = [\rho_1^N, \rho_2^N, \ldots, \rho_{\frac{m}{2}}^N, \frac{1}{2}\rho_\varnothing^N]$$

and note that the sum of these elements add up to 1.

We construct the first row of weights iteratively. For each $j \in \{1, \ldots, \frac{m}{2}+1\}$, let

$$w_{1,j} = \begin{cases} V_j^{N,1} & \text{if } V_1^E - \sum_{k=1}^{j-1} w_{1,k} \geq V_j^{N,1}, \\ V_1^E - \sum_{k=1}^{j-1} w_{1,k} & \text{otherwise.} \end{cases}$$

By the assumption of strict garbling $\rho_1^E \geq \rho_1^N$, and $w_{1,1} = \rho_1^N$. All other weights in the first row are either zero or positive with $V_1^E = \sum_{j=1}^{\frac{m}{2}+1} w_{1,j}$.

We now construct the rest of the weights row by row in an iterative process. At each step $i \in \{2, \ldots, \frac{m}{2}+1\}$, let

$$V^{N,i} = \left[\left(V_1^{N,1} - \sum_{k=1}^{i-1} w_{k,1}\right), \left(V_2^{N,1} - \sum_{k=1}^{i-1} w_{k,2}\right), \ldots, \left(V_{\frac{m}{2}+1}^{N,1} - \sum_{k=1}^{i-1} w_{k,\frac{m}{2}+1}\right)\right].$$

Iterating over $j \in \{1, \frac{m}{2}+1\}$, let

$$w_{i,j} = \begin{cases} V_j^{N,i} & \text{if } V_i^E - \sum_{k=1}^{j-1} w_{i,k} \geq V_j^{N,i}, \\ V_i^E - \sum_{k=1}^{j-1} w_{i,k} & \text{otherwise.} \end{cases}$$

By the assumption of strict garbling, $\sum_{k=1}^{j} w_{k,j} = V_j^{N,1}$. Thus, by the construction of the vector $V^{N,i}$, $w_{i,j} = 0$ for all $i > j$. Combined, these two conditions imply $\sum_i w_{i,j} = V_j^{N,1}$ for all j in the submatrix. Further, since both vectors V^E and $V^{N,1}$ sum to 1 by construction, $\sum_j w_{i,j} = V_i^E$ for all i.

Taken together, the construction of the submatrix generates a set of weights such that we can recover the elements of $V^{N,1}$ by adding the elements of the column together. As the first $\frac{m}{2}$ elements of $V^{N,1}$ correspond to $\{\rho_1^N, \ldots, \rho_{\frac{m}{2}}^N\}$, we can relate the weight matrix to ρ_j^N by adding the elements of column j together. Likewise, we can recover elements of V^E by adding the elements of the rows together. As the first $\frac{m}{2}$ elements of V^E correspond to $\{\rho_1^E, \ldots, \rho_{\frac{m}{2}}^E\}$, we can relate the weight matrix to ρ_i^E by adding the elements of row i together.

We now take advantage of symmetry to construct the weights for elements of W where $i \in \{\frac{m}{2}+1, \ldots, m+1\}$ and $j \in \{\frac{m}{2}+1, \ldots, m+1\}$. To avoid confusion with the previous step, let $k \in \{\frac{m}{2}+1, \ldots, m+1\}$ represent the rows in this submatrix of W and let $l \in \{\frac{m}{2}+1, \ldots, m+1\}$ represent columns. Next, let $w_{k,l} = w_{(m+2-k),(m+2-l)}$. Note that by reflection, $w_{k,l} = 0$ if $l > k$. All other weights are greater or equal to zero.

By symmetry, $\rho_k^E = \rho_{m+2-k}^E$. Thus for all $k \in \{\frac{m}{2}+2, \ldots, m+1\}$, $\rho_k^E = V_{m+2-k}^E$ and

$$\phi(s_k)\rho_k^E = \sum_{\{l | s_l \geq .5\}} \phi(s_k) w_{k,l}$$

Likewise, if k $= \frac{m}{2}+1$, $Q_{T\varnothing}^E = V_{m+2-k}^E$ and

$$\phi(s_\varnothing) Q_{T\varnothing} = \sum_{\{l | s_l \geq .5\}} \phi(s_\varnothing) w_{k,l}.$$

This implies

$$\left[\sum_{\{k|s_k>.5\}}\phi(s_k)\rho_k^E\right]+\phi(s_\varnothing)Q_{T\varnothing}^E=\sum_{\{k|s_k\geq.5\}}\sum_{\{l|s_l\geq.5\}}\phi(s_k)w_{k,l} \tag{29}$$

Using the same logic,

$$\phi(s_l)\rho_l^N=\sum_{\{k|s_k\geq.5\}}\phi(s_l)w_{k,l}$$

for all $l\in\{\frac{m}{2}+2,\ldots,m+1\}$ and

$$\phi(s_\varnothing)Q_{T\varnothing}^N=\sum_{\{k|s_k\geq.5\}}\phi(s_\varnothing)w_{k,l}.$$

when $l=\frac{m}{2}+1$. Thus

$$\left[\sum_{\{l|s_l>.5\}}\phi(s_l)\rho_l^N\right]+\phi(s_\varnothing)Q_{T\varnothing}^N=\sum_{\{k|s_k\geq.5\}}\sum_{\{l|s_l\geq.5\}}\phi(s_l)w_{k,l} \tag{30}$$

Subtracting (30) from (29) implies that (28) holds.

By Assumption 6, $\mathbb{E}V(\theta|T)<.75$. We now show that when $\mathbb{E}V(\theta|T)<.75$, $\phi(s_k)>\phi(s_l)$ if $s_k>s_l\geq.5$. By definition

$$\begin{aligned}\phi(s_k)-\phi(s_l)&=(2s_k-1)[s_k(1-\mathbb{E}V(\theta|T))+(1-s_k)\mathbb{E}V(\theta|T)]\\&\quad-(2s_l-1)[s_l(1-\mathbb{E}V(\theta|T))+(1-s_l)\mathbb{E}V(\theta|T)]\\&=2(s_k-s_l)[s_k(1-\mathbb{E}V(\theta|T))+(1-s_k)\mathbb{E}V(\theta|T)]\\&\quad-(2s_l-1)[(s_k-s_l)(2\mathbb{E}V(\theta|T)-1)]\\&=(s_k-s_l)[2s_k-4s_k\mathbb{E}V(\theta|T)+2V(\theta|T)-(2s_l-1)(2\mathbb{E}V(\theta|T)-1)]\end{aligned}$$

We have assumed that $s_k>s_l$. This implies that (s_k-s_l) is strictly positive and $\phi(s_k)>\phi(s_l)$ if and only

$$2s_k-4s_k\mathbb{E}V(\theta|T)+2\mathbb{E}V(\theta|T)-(2s_l-1)(2\mathbb{E}V(\theta|T)-1)>0. \tag{31}$$

Notice that (31) is decreasing in s_l. This implies that:

$$\begin{aligned}2s_k-4s_k\mathbb{E}V(\theta|T)+2\mathbb{E}V(\theta|T)-(2s_l-1)(2\mathbb{E}V(\theta|T)-1)\;>\;&2s_k-4s_k\mathbb{E}V(\theta|T)+2\mathbb{E}V(\theta|T)\\&-(2s_k-1)(2\mathbb{E}V(\theta|T)-1).\end{aligned}$$

Thus, a sufficient condition for $\phi(s_k)-\phi(s_l)$ to be positive is for

$$2s_k-4s_k\mathbb{E}V(\theta|T)+2\mathbb{E}V(\theta|T)-(2s_k-1)(2\mathbb{E}V(\theta|T)-1)\geq0.$$

Rearranging this equation, $\phi(s_k)-\phi(s_l)$ is positive if

$$\frac{\mathbb{E}V(\theta|T)-.25}{2\mathbb{E}V(\theta|T)-1}\geq s_k.$$

Further, noting that $s_k\in(.5,1]$, $\phi(s_k)-\phi(s_l)$ is positive if

$$\frac{\mathbb{E}V(\theta|T)-.25}{2\mathbb{E}V(\theta|T)-1}\geq1.$$

The LHS is decreasing in $\mathbb{E}V(\theta|T)$ and equal to one when $\mathbb{E}V(\theta|T)=.75$. Thus $\phi(s_k)>\phi(s_l)$ whenever $\mathbb{E}V(\theta|T)<.75$

By the construction of the weight matrix, there exists at least one element $w_{k,l}$ with $k>l$ where $w_{k,l}>0$. For this element $[\phi(s_k)-\phi(s_l)]w_{k,l}>0$. Noting that $w_{k,l}=0$ when $k<l$, it follows that all other elements of (28) are either positive or zero and thus

$$(\mathbb{E}V(Q^E|T)-\mathbb{E}M^V(\theta|Q^E,T))-(\mathbb{E}V(Q^N|T)-\mathbb{E}M^V(\theta|Q^N,T))$$

is positive.

Online Appendix D: Additional Empirical Results

D1. Estimated weights in Experiment 2

In this appendix, we estimate the weights of individual forecasters in Experiment 2. Our analysis is identical to that of Experiment 1 (see Section 3.1.2). For the SC algorithm, we once again estimate the prior by regressing the probability meta-prediction on the probability forecast and finding where this line crosses the identity line. Over this entire dataset for Experiment 2, the prior is estimated to be at .68. Thus, the data in this dataset is also biased towards true.

Figure 9 shows the estimated weights in the SP algorithm (top panel) and SC algorithm (bottom panel) as a function of the signal they received for all five grade levels combined. As before, the black solid line in each graph shows the predictions from each theoretical model while the dashed line shows the estimates from a non-parametric kernel regression. As seen in the top graph, both the magnitude of forecasters' signals ($|P_{i,k} - 0.5|$) and their votes ($V_{i,k}$) were significant negative predictors of their weight in the SP algorithm, $\beta_1 = -0.70$, $F(1, 458) = 2108.6$, $p < .001$ and $\beta_2 = -0.06$, $F(1, 458) = 177.0$, $p < .001$. There is once again a gap in the weight function at 0.5 in the same direction as before, which indicates that there is bias in the dataset towards answering true. However, unlike the states data, the gap is much smaller, and individuals who are certain that an event is true or false have weights that are less than half that of an individual who has a vote meta-prediction of 0.5.

As seen in the bottom panel, the SC algorithm has weights that are increasing in the distance away from the predicted prior, with a significant and large positive slope in the model specification that is consistent with the theoretical predictions, $\beta_1 = 0.22$, $F(1, 458) = 118.3$, $p < .001$. Better-informed forecasters are therefore generating larger weights in the SC algorithm than lesser-informed forecasters.

Our results here are therefore consistent with our theoretical model predictions: weights in the SP algorithm are decreasing in the distance from the 0.5 in the quiz dataset whereas weights in the SC algorithm are increasing in the distance away from the estimated uninformed prior.

D2. Weights and Expertise in Experiment 1 and 2

In the main text, we divided forecasters into high-performers and low-performers as a way of separating forecasters who are likely to be experts from those who are likely to be novices. In this section, we study an alternative specification where we further subdivide forecasters into quartiles to better understand how forecasters with different track-records contribute to the algorithm.

Similar to our approach in the main text, we sorted forecasters based on their mean accuracy on all other decision problems in the test set using leave-one-out cross-validation. Next, we divided forecasters into four quartiles containing equal numbers of forecasters and examined the average weight assigned by each algorithm to each quartile of forecasters over the test set. By construction, the weights in the four quartiles add up to one.

Figure 10 shows the alternative quartile specification for Experiment 1. As seen in the right hand panel of Figure 10, both the SP and SC algorithm over-weight forecasters in the highest two quartiles and under-weight forecasters in the lowest two quartiles on false questions. However, the SC algorithm assigns substantially

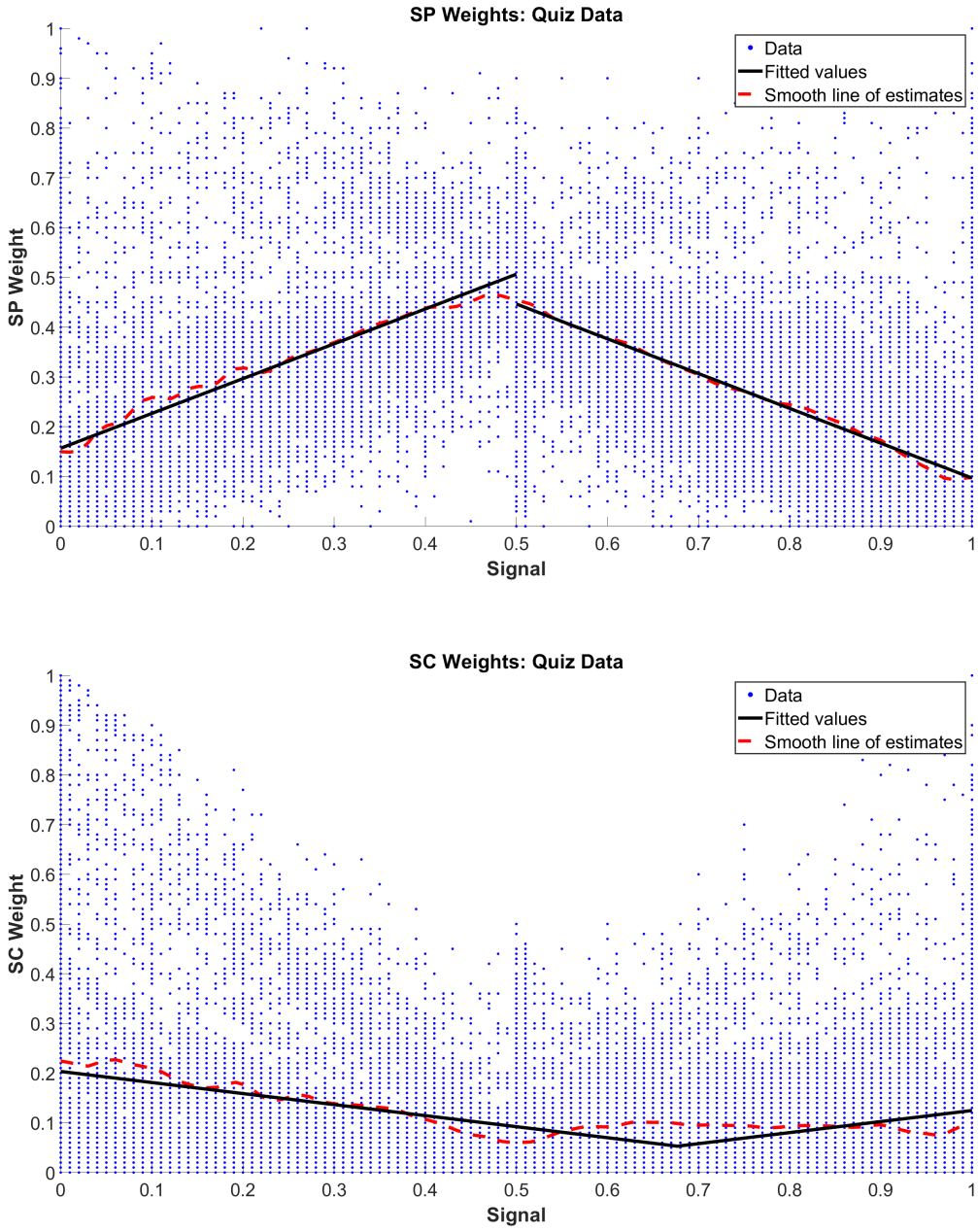


Figure 9 The relationship between forecasters' posterior and the weight assigned to them by the SP algorithm (top panel) and the SC algorithm (bottom panel) for the Quiz Data. The solid black lines are the predictions from the theoretical models. The dashed line is from a non-parametric kernel regression.

more weight to forecasters in the highest quartile and substantially less weight to forecasters in the lowest quartile, compared to the SP algorithm. As seen in the left panel, the SC also over-weights forecasters in the highest quartile and under-weights forecasters in the lowest quartile in true questions, while the weights in the SP algorithm are similar across the four quartiles. Thus, the SC algorithm appears to be more effective than the SP algorithm at assigning weight to forecasters who are correct more often on average.

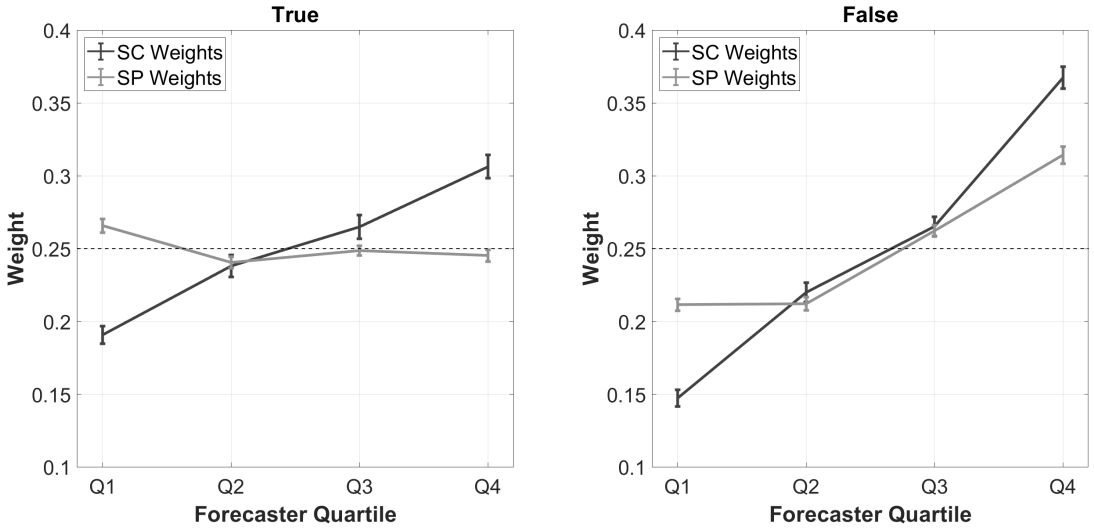


Figure 10 The average weight assigned by the SC algorithm and the SP algorithm as a function of forecasters' accuracy in the States dataset for the events where the outcome was "True" (left) and "False" (right). Forecasters were sorted into four bins, with 'Q1' containing the least accurate forecasters and 'Q4' containing the most accurate forecasters. The dotted line indicates the weights of an algorithm that weights all forecasters equally. Error bars represent the standard error.

We also applied the same analysis to the Quiz dataset from Experiment 2. We computed average weight assigned by the SC algorithm and SP algorithm assigned to each quartile of forecasters separately for each of the levels of difficulty and overall across all five levels of difficulty. As seen in Figure 11, both algorithms under-weight the least accurate forecasters and over-weight the most accurate forecasters in the more difficult datasets, but not necessarily in the easier dataset where a large proportion of forecasters are correct. On difficulty 2, the SC algorithm does a much better job at distinguishing between the best-performing and worst-performing individuals, whereas the SP algorithm assigns both groups approximately equal weights. Collapsing across all five difficulties (bottom right panel), the SC algorithm generates a larger aggregate weight for the most accurate forecasts and a smaller aggregate weight for the least accurate forecasters.

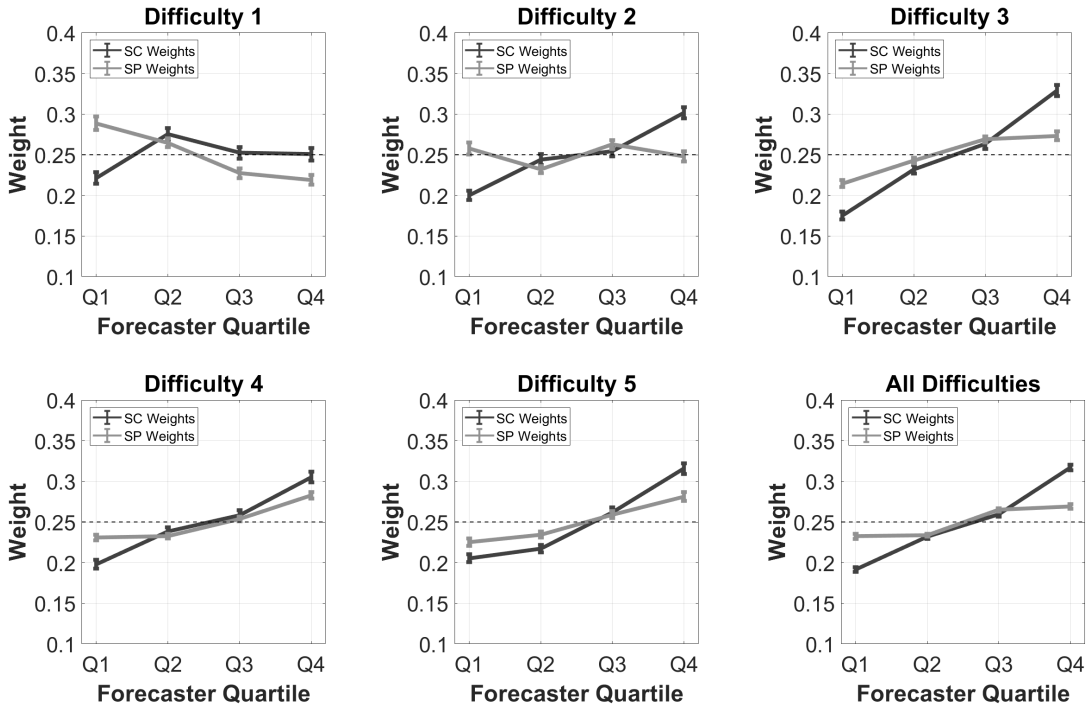


Figure 11 The average weight assigned by the SC algorithm and the SP algorithm as a function of forecasters' accuracy for each of the five individual difficulties and overall across all five difficulties in the Quiz dataset. Forecasters were sorted into four bins, with 'Q1' containing the least accurate forecasters and 'Q4' containing the most accurate forecasters. The dotted line indicates the weights of an algorithm that weights all forecasters equally. Error bars represent the standard error.

Online Appendix E: Forecaster Calibration Results

In this Appendix, we have included the calibration curves for individual forecasters' probabilistic forecasts in the States dataset from Experiment 1 and the Quiz dataset from Experiment 2. Figures 12 and 13 below shows the accuracy of individuals' forecasts with respect to their probabilistic forecasts with data binned by decile for each respective dataset. The relative size bubble for each decile shows the proportion of probability forecasts in the dataset in each decile. As can be seen in both figures, forecasters are overconfident when they would vote for true but not when they would vote for false. Thus, both the states data and the quiz data exhibit specific overconfidence in the framework discussed in Liberman and Tversky (1993) (also called overprediction in the parlance of Griffin and Brenner (2007)).

In both datasets, we found that the consensus vote was for true. Thus, the specific overconfidence observed is consistent with the consensus effect discussed in Koriat (2008) where forecasters who believe that the consensus answer is correct tend to be overconfident while forecasters who believe that the consensus answer is false tend to be underconfident.

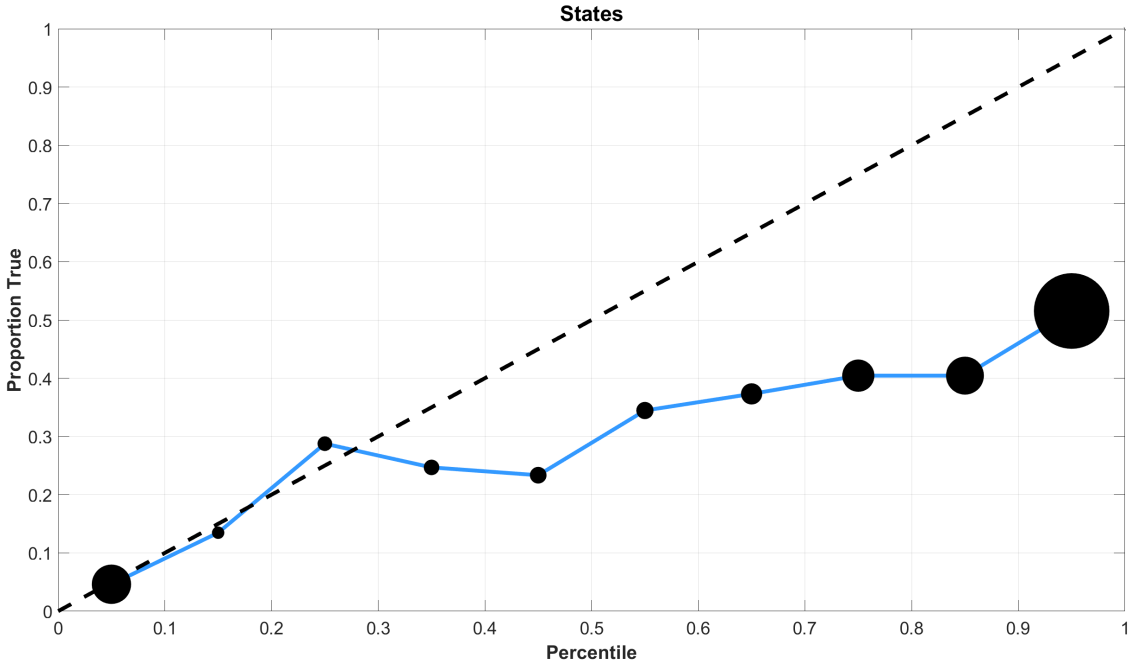


Figure 12 Calibration Curve for the States dataset showing the proportion of correct forecasts for each probability decile. The size of each bubble indicates the proportion of forecasts in that decile.

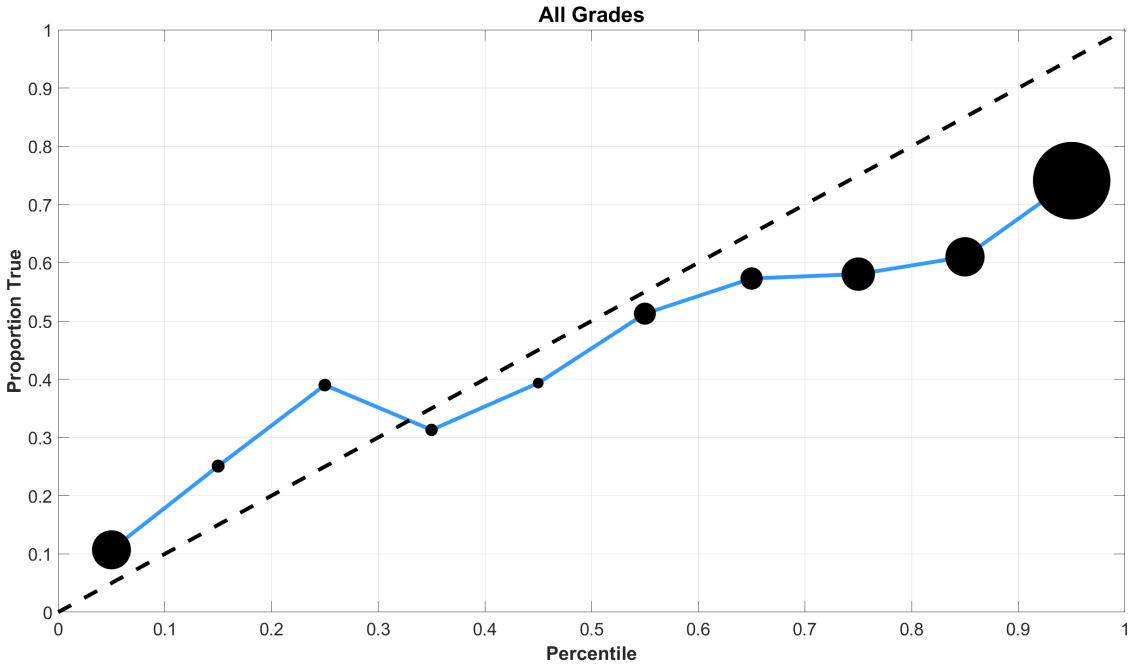


Figure 13 Calibration Curve for the Quiz dataset showing the proportion of correct forecasts for each probability decile. The size of each bubble indicates the proportion of forecasts in that decile.