BELIEF DECOMPOSITION
A MECHANISM FOR COLLECTIVE INFERENCE*

John McCoy† Dražen Prelec‡

Abstract

'Collective inference' is the task of computing the full information posterior of an event or proposition from expert judgments governed by incentives, where incentives depend only on these same judgments (no distributional assumptions or external verification). We formalize a Belief Decomposition mechanism to solve this task. The decomposition resolves beliefs into independent and shared components, which are then aggregated by Bayes' rule. The decomposition also enables incentives for probabilistic beliefs that reward independent, rather than shared, information. Unlike previous work, our mechanism applies to finite groups, including the minimal two-expert panel. Its empirical advantages are validated on three published datasets.

Key words: Proper scoring rules, Bayesian Truth Serum, Peer Prediction, Incentive-compatible surveys, Crowd wisdom, Information aggregation

JEL codes: C11, D82, D83, M00

I INTRODUCTION

Consider a set of experts with probabilistic beliefs about the 'state of the world.' An analyst would like to elicit these beliefs and combine them into a probability distribution, without imposing external priors on world states, expert information, or how these might be related. Two problems immediately arise from the perspective of mechanism design. The first is to define incentives for the honest expression of beliefs, with rewards linked to the informational value of revealed beliefs.

*We thank T. Wilkening, M. Marcellin, and P. Howe for the use of their data. Prelec thanks All Souls College, Oxford, for Visiting Fellowships during Michaelmas 2020, Hillary 2021, and Michaelmas 2022 terms.

†The Wharton School, University of Pennsylvania

‡MIT, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences; Email: dprelec@mit.edu

Second, the mechanism should compute the probability distribution consistent with all private and shared information in the sample.

This is a necessarily abstract statement of the problem, but concrete illustrations are easy to find. For example, here are three propositions posed to the Chicago-Booth panel of economic experts, for which experts answer on an Agree-Disagree scale, and indicate personal confidence: (a) "The typical chief executive officer of a publicly traded corporation in the U.S. is paid more than his or her marginal contribution to the firm's value." (b) "Not guaranteeing uninsured deposits at Silicon Valley Bank in full would have created substantial damage to the US economy." (c) "The average US citizen would be better off if a larger number of highly educated foreign workers were legally allowed to immigrate to the US each year."

These propositions are not verifiable, that is, they are not defined by contracts contingent on verifiable events. Example (a) uses imprecise language without an agreed upon method of resolution, (b) is a counterfactual, (c) is a prediction about a hypothetical policy that is unlikely to be implemented, and even if implemented would be difficult to assess.

Although not verifiable, all three propositions are understandable and their relevance to policy debates transparent. An assumption implicit in each question is that a true or best answer exists in principle, independent of policy preferences. In formal terms, there are two possible worlds: one where the claim, e.g., in question (c), is True and the average US citizen would be better off under the immigration scenario described in the question, and one where the claim is False. An expert panel may not have enough information to resolve the question, but by definition it has enough to assign a probability that best reflects all information distributed among its members. This observation applies to any form of 'crowd wisdom' prediction, such as forecasting the economic impact of technology and AI, or assessing foreign policy scenarios and catastrophic risks.

The classical approach (Condorcet, 1785; Galton, 1907) assumes that crowd wisdom can be extracted from the distribution of individual beliefs, provided those are free of bias. The Chicago-Booth survey is in this respect representative: the results are displayed as histograms on the Agree-Disagree scale, optionally weighted by confidence. The media then typically reports an average.[1] However, the distribution of individual beliefs does not determine the best collective probability distribution because essential detail about the information structure is missing (Arieli, Babichenko,

[1]Chicago-Booth is admittedly an informal platform, but the same methodology prevails in state-of-the-art technology forecasting. A representative example is a recent survey that estimated the impact of AI on labor, with respect to occupation, timing, and other variables (Grace et al., 2024). The survey was conducted on 2,778 AI researchers, and took care to neutralize biases and framing effects by rotating different question formats. However, the information request did not go beyond individual beliefs and estimates; results were then presented as distributions of estimates.

and Smorodinsky, 2020; Prelec, Seung, and McCoy, 2017, Proposition 1). In particular, individual beliefs do not reveal how information decomposes into shared and independent components. Without this decomposition, the distribution itself can be misleading. Indeed, the best collective probability distribution, i.e., the probability distribution implied by all information available to the panel, may point in a direction opposed to each individual's belief.

We define collective inference as the task of computing that best probability distribution from judgments governed by incentives, which have only these same judgments as inputs (no external, ground truth verification). We then construct a solution, *Belief Decomposition*, that elicits and incentivizes beliefs and predictions about the beliefs of others. The Belief Decomposition mechanism is introduced with a two expert example in the next section. A key new analytical device is an eigenvector-based decomposition of experts' beliefs and predictions into independent components, namely a common prior and a private signal for each individual. Computationally simple, this step appears to have been so far overlooked. Because the decomposition yields independent probabilities, the full information posterior is a straightforward application of Bayes' rule. The decomposition also facilitates truthtelling incentives that credit private information, distinct from the prior. The incentives are thus similar in spirit to market incentives, which reward information not reflected in the current price.

To inspire confidence and influence practice, a method for collective inference should deliver robust benefits in domains where accuracy can be easily verified. To that end, we assess performance on the three published data sets that have elicited the required inputs. Across these datasets, the posterior computed through belief decompostion improves on traditional methods by a substantial margin, with improvement increasing as problems become more challenging. For example, on a set of 100 science questions where average opinion is correct about 65% of the time, Belief Decomposition yields an accuracy rate of about 85%.

Predictions about others, usually about their 'type' rather than their probabilistic belief, have been used for mechanisms that incentivize honest non-verifiable answers (Prelec, 2004a; Witkowski and Parkes, 2012a; Radanovic and Faltings, 2013; Baillon, 2017; Cvitanić et al., 2019). Such mechanisms, however, do not address belief aggregation, and, except the Divergence-based truth serum (Radanovic and Faltings, 2014), do not apply to the minimal, two expert case.

Algorithms for aggregating individual judgments have also exploited predictions, starting with Prelec, Seung, and McCoy (2017), who proved that if experts vote for the world they find most likely, and predict the panel vote distribution, then in binary and certain other cases the true world

state will receive the most votes relative to predictions, in the large sample limit. The same contrarian principle is involved in several more recent algorithms (Palley and Soll, 2019; Wilkening, Martinie, and Howe, 2022; Palley and Satopää, 2023; Y.-C. Chen, Mueller-Frank, and Pai, 2021).

We discuss these incentives and aggregation literatures in Section VII. Although there are certain differences in model assumptions, the aggregation results are generally asymptotic, proving that the correct world state will be identified given a large, strictly speaking infinite, expert sample. In this advantageous situation the problem of constructing incentives without verification also dissolves, because the computed world state can proxy for verified ground truth (Y.-C. Chen, Mueller-Frank, and Pai, 2021). But, as the phrase 'second opinion' suggests, available experts are often few in number. This sets the stage for the motivating example in the next section, and the paper that follows.

II A TWO EXPERT EXAMPLE

Imagine a research proposal that is assessed on a stand-alone basis, apart from other competition. To evaluate it, a review panel recruits two experts, A and B. In keeping with standard practice for remote evaluation, these experts are anonymous and submit their judgements without communication.

The panel intends to fund the proposal if the chance of 'long term high-impact' results is greater than 50%. Experts are therefore asked to think in terms of only two possible outcomes for funding, $\omega_1 = \textit{Success}$, that the proposal will have long-term high-impact, and $\omega_2 = \textit{Failure}$, that it will not have such impact. This is severely stylized, but will illustrate the main ideas. Experts convey their assessment as a subjective probability of *Success*. They also predict the probability of *Success* provided by the other expert. Let us assume that these judgments are correct given the information available to each expert. Here are concrete numbers for illustration:

- expert A assesses a *Success* probability $x_1^A = 40\%$ and predicts $y_1^A = 30\%$ for B
- expert B assesses a *Success* probability $x_1^B = 25\%$ and predicts $y_1^B = 20\%$ for A

Before reading on, the reader might pause and guess what probability on *Success* is implied by these assessments.

If we simply compare experts' Success probabilities against the 50% threshold, the panel should decline funding as both fall short of 50%. For that matter, so do their predictions of each other's

probabilities. Nevertheless, a negative panel decision is likely a mistake here. A simple Bayesian interpretation of these four values implies a 67% chance of Success, well above threshold. Figure I summarizes the situation. The argument is elementary in the sense that it mostly involves applications of Bayes rule..

By 'simple Bayesian interpretation,' we mean a model that assumes a common prior and independent identically distributed signals. From the perspective of such a model, a striking aspect of the four numbers is that each expert believes that he or she is more more optimistic than their colleague. This pattern suggests that there is information available to each expert that is more optimistic than the evidence which is common knowledge and represented through the common prior. Common knowledge in this example includes the applicant's CV. Both experts can see, for example, that the author is an unknown scholar with a thin publication record, and assume that the other expert will assess these facts in the same way. Based on this common knowledge, the proposal should not be funded. However, each expert regards different aspects of the novel idea as promising, albeit not enough to move their individual assessment above 50%.

The probability assessed by expert A is the Bayesian posterior conditional on their received signal s^A and knowledge of the world prior $p(\omega_1)$, and similarly for expert B (given signal s^B):

$$x_1^A = p(\omega_1|s^A) \propto p(s^A|\omega_1)p(\omega_1)$$
$$x_1^B = p(\omega_1|s^B) \propto p(s^B|\omega_1)p(\omega_1)$$

Under the *i.i.d* signal model, the Bayesian posterior probability of $\omega_1 = Success$ is given by:[2]

$$\begin{aligned} p(\omega_1|s^A,s^B) &\propto p(s^A,s^B|\omega_1)p(\omega_1) \\ &\propto \frac{x_1^a \times x_1^B}{p(\omega_1)} = \frac{(.40)\times(.25)}{p(\omega_1)} \end{aligned}$$

where $p(\omega_1|s^A)$ and $p(\omega_1|s^B)$ are the subjective probabilities that experts A and B assign to $\omega_1 = Success$, given signals s^A and s^B, and $p(\omega_1)$ is the common prior probability of ω_1, i.e., the probability consistent with common knowledge. Because each expert combines the common prior with his independent signal, a simple product of experts' subjective probabilities would give the prior too much weight, double-counting it in effect. The Bayesian posterior discounts the product

[2]By Bayes' rule and conditional independence, $p(\omega_1|s^A,s^B) = p(s^A,s^B|\omega_1)p(\omega_1)/p(s^A,s^B) = [p(\omega_1|s^A)p(\omega_1|s^B)/p(\omega_1)] \times [p(s^A)p(s^B)/p(s^A,s^B)]$.

of declared probabilities by the common prior, which is so far unspecified.

To derive the common prior from the four inputs, we first write predictions as a weighted average of expected probabilities conditional on the world state, with weights given by the probabilities that each expert assigns to each world state:

$$\begin{aligned} E[x_1^B|s^A] =& p(\omega_1|s^A)E[x_1^B|\omega_1,s^A] + p(\omega_2|s^A)E[x_1^B|\omega_2,s^A] \\ E[x_1^A|s^B] =& p(\omega_1|s^B)E[x_1^A|\omega_1,s^B] + p(\omega_2|s^B)E[x_1^A|\omega_2,s^B] \end{aligned} \tag{1}$$

$E[x^B|\omega_j,s^A]$ is A's expectation of the probability that B will assess for $\omega_1 = Success$ on the hypothesis that the world is ω_j. Under the assumption that signals are i.i.d. given the world state, such conditional expectations will be the same for both experts, so we can express them as entries of a 2×2 stochastic matrix W:

$$\begin{aligned} E[x_j^B|\omega_i,s^A] =& E[p(\omega_j|s^B)|\omega_i,s^A] \\ =& E[p(\omega_j|s^B)|\omega_i] \\ =& w_{ij} \end{aligned} \tag{2}$$

The first line relies on honesty, the second on conditional independence, and the third on exchangeability (the identically distributed part in i.i.d.), which lets us drop the label B.

We call $W = [w_{ij}]$ the *local expectations* matrix, as it contains expectations about others' beliefs 'local' to each possible world state. It reduces a pair of conditional belief densities, one for each world, to two free parameters, w_{11} and w_{21}, which, if predictions are honest, can be derived from eq. 1, by solving:

$$\begin{aligned} .30 &= .40w_{11} + .60w_{21} \\ .20 &= .25w_{11} + .75w_{21} \end{aligned} \tag{3}$$

giving $w_{11} = 7/10$ and $w_{21} = 1/30$.

Eq. 1 should also hold for an uninformative signal, i.e. when $p(\omega_1|s^A) = p(\omega_1)$. An uninformative signal results in a prediction $E[x_1^B|s^A] = p(\omega_1)$. From that special case we derive the prior, $p(\omega_1) = p(\omega_1)w_{11} + (1-p(\omega_1))w_{21}$, or:

$$p(\omega_1) = \frac{w_{21}}{w_{21}+(1-w_{11})} = \frac{(1/30)}{(1/30)+(3/10)} = .10 \tag{4}$$

Given the low prior on *Success* of 10% for this proposal, the experts' beliefs of 40% and 25% indicate positive independent information, which when taken into account, imply 2:1 odds in favor

of $\omega_1 = \textit{Success}$ (see Figure I, caption). More generally, the stationary distribution of W, which can be obtained via an eigenvector decomposition, provides the prior.

So far, the example assumes that beliefs and predictions are honest, which might be open to question here, and in general. The accuracy of experts' predictions about the beliefs of others is verifiable, hence predictions can be scored with incentive-compatible proper scoring rules (Gneiting and Raftery, 2007). Incentives for unverifiable beliefs require something new.

A simple solution for two individuals, assuming i.i.d. signals, is to give both experts a fixed bonus if their beliefs are different and also strictly aligned with their predictions, so that the more optimistic expert predicts a higher average belief, $x_1^A > x_1^B \iff y_1^A > y_1^B$. This could be implemented, for example, with quadratic scoring of predictions:

$$V^A(x_1^A, y_1^A, x_1^B, y_1^B) = \begin{cases} -(y_1^A - x_1^B)^2 + 1 & if\ (x_1^A - x_1^B)(y_1^A - y_1^B) > 0 \\ -(y_1^A - x_1^B)^2 & otherwise \end{cases} \tag{5}$$

for expert A, and similarly for V^B. The quadratic penalty for predictions $(y_1^A - x_1^B)^2$ is minimized by setting $y_1^A = E[x_1^B | s^A]$ (this is the Brier score, Cooke (1991)). Regarding x_1^A, if signals have a continuous distribution with full support, any deviation from honest beliefs risks misalignment. That is, if A misreports their belief as $x \neq p(\omega_1 | s^A)$, then, by full support, there is some chance that the signal B received will result in $(x - x_1^B)(y_1^A - y_1^B) < 0$. Although the truthtelling equilibrium with eq. 5 is not unique, there is no strict equilibrium in which experts ignore their signals. Colluding to report identical beliefs triggers the penalty as $(x_1^A - x_1^B)(y_1^A - y_1^B) = 0$. Colluding to report different beliefs would lead to misalignment because each expert's predictions would be identical to the other expert's beliefs, $y_1^B = x_1^A$, $y_1^A = x_1^B$ implies $y_1^A < y_1^B \iff x_1^A > x_1^B$. The mechanism therefore compels attention to signals.

The design space for incentives expands if there are more than two experts. We will take advantage of this to credit the provision of information according to Shannon information gain. The main idea for incentives can be explained with three experts A, B and C. If the question is binary, then the inputs of any pair are sufficient to derive W and the prior. This means that Expert A can safely assume that the mechanism will compute correct w_{ij} values using the inputs of B and C only. Treating these values as fixed, eq. 1 generates an *implied prediction*, $\hat{y}_1^A = x_1^A w_{11} + (1 - x_1^A) w_{21}$, for A, as a function of the reported beliefs x_1^A. Incentives for x_1^A can be set by scoring $\hat{y}_1^A$ against x_1^B or x_1^C.

III THE SETUP

III.A Possible world model

The object of interest is the unknown and unverifiable world state $\omega_i \in \{\omega_1, \ldots, \omega_m\}$, a realization of the random variable Ω. There are n experts, labeled $r = 1, \ldots, n$. Each expert r receives a signal $s^r = (s^r_1, \ldots, s^r_m) \in \Delta(\Omega)$, a realization of the random variable S^r. The vector $S = (S^1, \ldots, S^n)$ denotes the random signal vector across all experts, with realization $s = (s^1, \ldots, s^n)$.

When the random variable is clear from the context, we may denote the probability of it taking a particular realization without writing the random variable, for example we may write $p(\Omega = \omega_i)$ as $p(\omega_i)$, $p(S^r = s^r)$ as $p(s^r)$ or, in the case of conditional distributions, $p(S^r = s^r | \Omega = \omega_j)$ as $p(s^r|\omega_j)$.

We begin by defining a *possible world model*. We impose five conditions that it should satisfy, all of which are used in the literature.

Definition 1. A *possible world model* $p = p(s, \omega_i) \in \mathscr{P}$ is a joint distribution over world states and signal vectors. Such a joint distribution can be decomposed into a prior over world states $p(\omega_i) = \Pr(\Omega = \omega_i)$ and a conditional distribution $p(s|\omega_i) = p(S^1 = s^1, \ldots, S^n = s^n | \Omega = \omega_i)$ for each world i.

Assumption 1. *The possible world model p is common knowledge among experts. The analyst only knows* $\mathscr{P}$.

Since experts do not predict the predictions of others, in principle we need only the assumption that each expert knows the possible world model, and knows that every other expert knows the possible world model.

Assumption 2. *Conditional independence. For any world state* ω_i,

$$p(s, \omega_i) = \prod_{r=1}^{n} p(s^r|\omega_i) p(\omega_i)$$

Conditional independence is a substantive constraint implying that the total amount of independent evidence scales with sample size, as each additional expert contributes something new. We will discuss in Section VIII, how our method can be modified if Assumption 2 is too strong.

Definition 2. Full information posterior: $p(\omega_i|s) = p(\Omega = \omega_i | S^1 = s^1, \ldots, S^n = s^n)$.

Under Assumptions 1 and 2, the full information posterior on ω_i is the normalized product of n expert probabilities on ω_i 'discounted' $n-1$ times by prior probability, $p(\omega_i)$:

$$\begin{aligned} p(\omega_i|s) =& \frac{p(\omega_i)}{p(s)} p(s|\omega_i) \\ =& \frac{p(\omega_i)}{p(s)} \prod_{r=1}^{n} p(s^r|\omega_i) \\ =& \frac{p(\omega_i)}{p(s)} \prod_{r=1}^{n} (p(\omega_i|s^r) \frac{p(s^r)}{p(\omega_i)}) \\ \propto& p(\omega_i)^{1-n} \prod_{r=1}^{n} p(\omega_i|s^r) \end{aligned}$$

The full information posterior corrects for the overweighting of the common prior in the product of individual beliefs. Each of the n experts incorporates the prior into their beliefs, hence the product double-counts the prior exactly $n-1$ times.

Assumption 3. *Exchangeability. For any two experts r, r', world state ω_i, and $q \in \Delta(\Omega)$*

$$p(S^r = q|\Omega = \omega_i) = p(S^{r'} = q|\Omega = \omega_i)$$

By exchangeability, each possible world model p generates a unique $m \times m$ row stochastic local expectations matrix W,

$$W = [w_{ij}] = [\int_{s^r} p(\Omega = \omega_j|S^r = s^r) p(S^r = s^r|\Omega = \omega_i)] \tag{6}$$

which does not depend on the selection of expert r. The matrix W will be key to our formulation of incentives and aggregation. As already noted, the stationary distribution of W is the world prior, $p(\Omega = \omega_j)$.

The final pair of assumptions guarantee, respectively, the computability and invertibility of W.

Assumption 4. *Full support common prior. For all experts r, world states ω_i, and $s^r \in \Delta(\Omega)$: $p(s^r, \omega_i) > 0$.*

Under full support, a random sample of n experts will generically generate n distinct, linearly independent signals. This will ensure that the row stochastic matrix W can be derived by linear regression of belief predictions on beliefs.

Assumption 5. *Stochastic relevance. For all experts r, t, and $q, q' \in \Delta(\Omega)$, $q \neq q' \implies E[S^t|S^r =$*

$q] \neq E[S^t|S^r = q']$.

In our setting, this assumption is needed for incentives but not for aggregation. It states that differences in world-state beliefs imply differences in expectations of another expert's beliefs, which implies in turn that W has full rank.

III.B Beliefs do not constrain the full information posterior

Section II contained an example where the full information posterior did not lie between the two individual estimates. It is thus natural to ask whether beliefs place any constraint on the full information posterior. This question has been raised and answered previously, essentially in the negative (Arieli, Babichenko, and Smorodinsky, 2020; Prelec, Seung, and McCoy, 2017), but is worth addressing directly in our setting.

Lemma 1. *Fix n belief vectors $p(\omega|s^1), \ldots, p(\omega|s^n)$ over $m \leq n$ possible worlds. Then for any probability distribution over worlds, $(q_1, \ldots, q_m)$, $q_k > 0$, there exists a prior over worlds that induces $(q_1, \ldots, q_m)$ as the full information posterior.*

Proof. Under full support, a finite sample of signals places no constraints on the world prior. By setting the world prior as:

$$\tilde{p}(\omega_i) = \left(\frac{q_i}{\prod_{r=1}^{n} p(\omega_i|s^r)}\right)^{1/(1-n)}$$

and substituting the prior into the expression for the full information posterior (Definition 2), we obtain the desired distribution $q_i = p(\omega_i|s)$. ■

III.C Collective inference mechanism

Experts feed the collective inference mechanism by submitting reports, whose specific format we will shortly define. From these reports an uninformed analyst, i.e., a machine, should be able derive the full information posterior in Definition 2. Because experts' first-order beliefs alone do not determine the posterior (Lemma 1), the reports need to include further details about the possible world model p. This additional information might be higher-order belief predictions, hypothetical beliefs conditional on world states, or, in a brute force approach, the complete possible world model. Choosing among these options involves both theoretical and practical considerations, such as whether experts will understand the information request and whether elicitation is too demanding. The definition below accommodates different approaches.

Denote by $\mathscr{A}$ the space of possible reports (inputs), $a^r \in \mathscr{A}$ the report by expert r, a the vector of all reports, and a^{-r} the vector of reports excluding r. A mechanism $\mathscr{M} = \langle \mathscr{A}, (V^r)_{r=1}^n, f \rangle$ consists of $\mathscr{A}$, an ensemble of real-valued scoring functions, $V^r : \mathscr{A} \to \mathbb{R}$, $r = 1, \ldots, n$, and an inference function $f : \mathscr{A} \to \Delta(\Omega)$ that computes a probability distribution over world states.

Let $\sigma^r : S^r \to \mathscr{A}$, denote a pure strategy of expert r, $\sigma(s) = (\sigma^1(s^1), \ldots, \sigma^n(s^n))$ the strategy profile across all experts, and $\sigma^{-r}(s^{-r})$ the profile excluding r. A strategy profile μ is a *signal-pooling strategy profile* if $\mu(s) = a = (a^1, \ldots, a^n) = \mu(s')$ for any signal vectors s, s'. Note that this definition of pooling does not imply that experts all give the same report, only that they ignore their signals.

Definition 3. $\mathscr{M} = \langle \mathscr{A}, (V^r)_{r=1}^n, f \rangle$ is a collective inference mechanism on $\mathscr{P}$ if for all possible world models $p \in \mathscr{P}$ there exists a strategy profile σ_p such that (1) and (2) hold. It is pooling-proof if (1)-(3) hold.

1. σ_p is a strict Bayesian Nash equilibrium:

$$a^r \neq \sigma_p^r(s^r) \implies E[V^r(\sigma_p^r(s^r); \sigma_p^{-r}(s^{-r}))|S^r = s^r] > E[V^r(a^r; \sigma_p^{-r}(s^{-r}))|S^r = s^r]$$

2. the inference function applied to $\sigma_p(s)$ computes the full information posterior:

$$f(\sigma_p(s)) = (p(\omega_1|s)), \ldots, p(\omega_m|s))$$

   or, for some, but not all signals, yields a null result, $f(\sigma_p(s)) = \emptyset$.

3. any Bayesian Nash equilibrium μ defined by signal-pooling strategies, $\mu(s) = a = (a^1, \ldots, a^n)$ is dominated by σ_p in expected score:

$$E[V^r(\sigma_p(s))|S^r = s^r] > V^r(a),\ r = 1, \ldots, n$$

The expectation of the score of expert r is taken with respect to the conditional distribution, $p(s^{-r}|s^r)$ over reports of other experts in equilibrium, $a^{-r} = \sigma_p^{-r}(s^{-r})$. The mechanism should never mislead with an incorrect posterior; however, it may admit defeat and fail to compute a posterior. This may happen out of equilibrium, or if signal combinations contain insufficient information about the possible world model.

III.D Proper scoring rules

Mechanism incentives will make use of proper scoring rules. Let θ be a realization in Δ^{m-1} of random vector Θ, and $\hat{\theta}$ a prediction of θ based on some joint distribution $q(\theta,\psi) = \Pr(\Theta = \theta, \Psi = \psi)$. We refer to θ as the 'scoring target' of prediction $\hat{\theta}$, or to $\hat{\theta}$ being 'scored against' θ. A (real-valued) scoring rule $u(\theta,\hat{\theta})$ is strictly proper if $E_q[u(\theta,\hat{\theta})|\Psi = \psi]$ is uniquely maximized for $\hat{\theta} = E_q[\theta|\Psi = \psi]$. If $u(\theta,\theta) = 0$, and $\bar{u}_{min} \equiv \min_q E_q[u(\theta, E_q[\theta])]$ is finite, then it is a *strictly proper penalty*. Two examples of strictly proper penalties are the negative K-L divergence and the quadratic Brier score:

$$u(\theta,\hat{\theta}) = \sum_{k=1}^{m} \theta_k \log \frac{\hat{\theta}_k}{\theta_k} \leq 0, \qquad \bar{u}_{min} = -\log m$$

$$u(\theta,\hat{\theta}) = -\sum_{k=1}^{m} (\theta_k - \hat{\theta}_k)^2 \leq 0, \;\; \bar{u}_{min} = 1/m - 1$$

Both scoring rules are symmetric, penalizing equally errors on all coordinates (Gneiting and Raftery, 2007). The maximum expected penalty for a symmetric rule is obtained if ψ is uninformative, and $q(\theta|\psi) = q(\theta)$ is a uniform distribution over the corners of Δ^{m-1}, i.e., $\theta_k \in \{0,1\}$, $\sum_k \theta_k = 1$, $\hat{\theta}_k = 1/m$.

IV BELIEF DECOMPOSITION

The BD mechanism generalizes the example in Section II to n experts. We present two versions, one that requires strictly more experts than world states, and one that also covers the $n = m$ case and generalizes the scoring proposal in Section II. In Section V, we also describe an approximate mechanism for when there are fewer experts than worlds, and show via simulations that the approximation error is likely small.

IV.A Belief Decomposition for more experts than worlds

Definition 4. Belief decomposition mechanism $\mathscr{M} = < \mathscr{A}, (V^r)_{r=1}^n, f >$ for $n > m$.

1. Reports. Each expert r submits a pair of probability vectors, $a^r = (x^r, y^r) \in \mathscr{A} = \Delta(\Omega) \times \Delta(\Omega)$, with x^r interpreted as beliefs over worlds, and y^r as prediction of: $\bar{x}^{-r} = (n-1)^{-1} \sum_{r' \neq r} x^{r'}$, the average of other experts' beliefs.

2. Local expectations. Combine all (x^r, y^r) into $n \times m$ row-stochastic matrices $X \equiv [x^1, \ldots, x^n]$, $Y \equiv [y^1, \ldots, y^n]$, and estimate *the* $(m \times m)$ *local expectations matrix* W:

$$\tilde{W} = (X^T X)^{-1} X^T Y \tag{7}$$

Similarly, for each r combine reports excluding r into $(n-1) \times m$ matrices X^{-r}, Y^{-r}, and estimate $\tilde{W}^{-r}$.

3. Computability trigger, $C \in \{0,1\}$: If $\tilde{W}$ or any $\tilde{W}^{-r}$ is is not a computable row stochastic matrix or is the identity matrix, then set $C = 0$. Otherwise, set $C = 1$.

4. Prior. If $C = 1$, estimate the world prior as the left unit eigenvector $\tilde{p}(\omega) = (\tilde{p}(\omega_1), \ldots, \tilde{p}(\omega_m)) \in \Delta(\Omega)$ of $\tilde{W}$:

$$\tilde{p}(\omega)\tilde{W} = \tilde{p}(\omega) \tag{8}$$

Similarly, for each r estimate $\tilde{p}(\omega)^{-r} \in \Delta(\Omega)$ as the left unit eigenvector of $\tilde{W}^{-r}$.

5. Implied predictions. If $C = 1$, compute $\hat{y}^r \in \Delta(\Omega)$, which is a function of x^r:

$$\hat{y}_i{}^r(x, y^{-r}) = \sum_k x_k^r \tilde{w}_{ki}^{-r}, \;\; i = 1, \ldots, m \tag{9}$$

6. Scoring. Select a strictly proper penalty u, parameters $\alpha \in (0,1)$, $\lambda \geq 0$, and assign a score to each r:

$$V^r(x,y) = \begin{cases} \alpha u(\bar{x}^{-r}, y^r) + (1-\alpha) u(\bar{x}^{-r}, \hat{y}^r) + \lambda u(y^r, \hat{y}^r) - u(\bar{x}^{-r}, \tilde{p}(\omega)^{-r}) & if\ C = 1 \\ u(\bar{x}^{-r}, y^r) & if\ C = 0 \end{cases} \tag{10}$$

7. Inference function, $f(x,y) \in \Delta(\Omega)$. Compute the full information posterior if possible:

$$f_i(x^1, \ldots, x^n, y^1, \ldots, y^n) \begin{cases} \propto \prod_{r=1}^n x_i^r \tilde{p}(\omega_i)^{1-n} & if\ C = 1 \\ = \emptyset & if\ C = 0 \end{cases} \tag{11}$$

Proposition 1. *If $\mathscr{P}$ satisfies Assumptions 1-5, the number of experts exceeds the number of worlds $(n > m)$, and $\mathscr{M} = \langle \mathscr{A}, (V^r)_{r=1}^n, f \rangle$ is constructed according to Definition 4, then $\mathscr{M}$ is a pooling-*

proof collective inference mechanism on $\mathscr{P}$.

Appendix I provides the proof. The local expectation matrix is estimated first (as $\tilde{W}$) by linear regression of Y on X (eq. 7). The assumption of full support is critical here, ensuring that a random sample of signals will generically yield a matrix X with m linearly independent columns. (This computational step also figures in the PMBA method of Y.-C. Chen, Mueller-Frank, and Pai (2021), and, under a slightly different setup, in Libgober (2021).)

The stationary distribution (i.e., the unit eigenvector) of $\tilde{W}$ then provides an estimate of the world prior (eq. 8). The inference function f is a straightforward application of Bayes' rule (eq. 11).

The scoring function in 10 consists of four terms: prediction accuracy $\alpha u(\bar{x}^{-r}, y^r)$, implied prediction accuracy $(1-\alpha)u(\bar{x}^{-r}, \hat{y}^r)$, a prediction garbling score $\lambda u(y^r, \hat{y}^r)$, and a side-payment $-u(\bar{x}^{-r}, \tilde{p}(\omega)^{-r})$. The first of these, prediction accuracy, incentivizes honest belief predictions by proper scoring. The scoring target for belief predictions y^r are beliefs averaged across the sample, excluding r from the average.

As discussed in Section II, scoring beliefs presents a challenge: the world state is not verifiable and the full information posterior over worlds is not a suitable proxy target as it depends on x^r in a finite sample. If, however, the full information posterior is computed without x^r, then its expectation conditional on s^r would not match s^r. Our solution is based on the observation that an honest equilibrium implies $W = W^{-r}$, and once W^{-r} is specified, beliefs over worlds x^r generate an implied prediction $\hat{y}^r(x^r)$ that should be the same as y^r. Because W^{-r} does not involve x^r or y^r, one can treat $\bar{x}^{-r}$ as the target for proper scoring of $\hat{y}^r$. The prediction garbling term is an optional penalty on deviations between y^r and $\hat{y}^r$. These redundancies in the scoring function are included for robustness and flexibility. Empirical results will clarify the optimal setting of weights α, λ, and Section VI provides some initial evidence on this point.

An interesting alternative, which we do not formally incorporate, would be to invert the (square) matrix $\tilde{W}^{-r}$ and treat *implied beliefs* $\hat{x}^r$ as the scoring target for beliefs, with the former derived by applying $(\tilde{W}^{-r})^{-1}$ to y^r. This would score beliefs directly, in effect telling experts what their beliefs 'should have been' given their predictions. Under this regime, an expert would receive feedback in the form of a direct error signal on stated beliefs.

The final side-payment term provides participation incentives. As $y^r_k = \hat{y}^r_k$ in an honest equilibrium σ, we have $E_\sigma[u(y^r_k, \hat{y}^r_k)|s^r] = 0$, and

$$E_\sigma[V^r(x,y)|s^r] = \sum_k E_\sigma[u(\bar{x}_k^{-r}, y_k^r)|s^r] - \sum_k E_\sigma[u(\bar{x}_k^{-r}, \tilde{p}(\omega_k)^{-r})|s^r]$$
$$> 0$$

The inequality is strict because u is a strict penalty and $y_k^r = E_\sigma[\bar{x}_k^{-r}|s^r] \neq \tilde{p}(\omega_k)^{-r} = p(\omega_k)$ (except on a set measure zero). In the case when scoring is logarithmic, $u = \log$:

$$E_\sigma[V^r(x,y)|s^r] = \sum_k E_\sigma[\bar{x}_k^{-r}|s^r] \log \frac{y_k^r}{\tilde{p}^{-r}(\omega_k)}$$
$$= \sum_k y_k^r \log \frac{y_k^r}{p(\omega_k)},$$

expected scores in an honest equilibrium correspond to Shannon information gain computed when prior beliefs are replaced by belief predictions. The expected score for expert r measures how much information about other experts' beliefs is contained in signal s^r. To the extent that experts can influence the possible world model by, for example, gathering information in advance of the elicitation process, they will have an incentive to search for evidence that is both relevant to possible world identification and outside of common knowledge. Conversely, shifting to a garbled model, i.e., one where signals are corrupted by noise, would reduce experts' expected scores.

IV.B Belief Decomposition when $n = m$

Belief Decomposition incentives as given by eq. 10 do not work for $n \leq m$ because X^{-r} has fewer rows than columns and $\tilde{W}^{-r}$ cannot be computed as in eq. 7. To cover $n = m \geq 2$, we generalize the scoring proposal in Section II, which assigns a fixed bonus if beliefs and predictions are aligned (eq. 5). To extend this to more than two worlds, the bonus is conditioned on the computability of W as a row stochastic matrix. One can show that reporting incorrect beliefs by a single expert when others experts are honest creates a non-zero risk of failing this condition and losing the bonus. The risk arises because another expert might a receive a similar signal, in which case her predictions will likely be close to the dishonest expert's predictions, but the stated beliefs may differ by a larger amount. As with the example in Section II, the argument relies on full support (Assumption 4).

Relative to Definition 4, the maximum expected penalty, $\bar{u}_{min}$, replaces the side-payment $-u(\bar{x}^{-r}, \tilde{p}(\omega)^{-r})$.

Definition 5. Belief Decomposition mechanism $\mathscr{M} = < \mathscr{A}, (V^r)_{r=1}^n, f >$ for $n \geq m$. Follow Definition 4 Steps 1-7, with the following modification to Step 6:

6. Select a strictly proper penalty u and assign a score to each r:

$$V^r(x,y) = \begin{cases} u(\bar{x}^{-r}, y^r) - \bar{u}_{min} & if\ C = 1 \\ u(\bar{x}^{-r}, y^r) & if\ C = 0 \end{cases} \tag{12}$$

Proposition 2. *If $\mathscr{P}$ satisfies Assumptions 1-5, the number of experts is greater or equal to the number of worlds ($n \geq m$), and $\mathscr{M} = \langle \mathscr{A}, (V^r)_{r=1}^n, f \rangle$ is constructed according to Definition 5, then $\mathscr{M}$ is a pooling-proof collective inference mechanism on $\mathscr{P}$.*

The proof (in Appendix I) exploits the fact, noted above, that predictions of different experts should be more similar than their beliefs, since predictions are a garbling of beliefs through matrix W. For example, in the dyad case a pair of reports (x^A, y^A), (x^B, y^B) will produce a stochastic matrix W if and only if:

$$\begin{aligned} x_1^A > x_1^B &\iff y_1^A > y_1^B \\ x_1^A > x_1^B &\iff \frac{x_1^A}{x_1^B} > \frac{y_1^A}{y_1^B} > \frac{1-x_1^A}{1-x_1^B} \end{aligned} \tag{13}$$

In addition to alignment (top line), honesty also implies that the ratio of beliefs is more extreme than the ratio of predictions (bottom line). In the general case, 'more similar' may be defined by quadratic distance.

V BELIEF DECOMPOSITION UNDER WEAKER ASSUMPTIONS

The principle that aggregation should proceed over independent pieces of evidence, rather than individual beliefs, remains valid even if our particular model assumptions do not hold. It is useful to know whether assumption failure requires significant modification of the theory or whether one can proceed with some relatively simple adjustment, as described in the three examples below.

V.A BD with sparse data, $n < m$

Applying BD to continuous variables, for example to predict a continuous quantity, is straightforward in principle; the continuous variable is partitioned into a finite number of bins. The number of bins, however, sets a lower limit on expert panel size, which could be large if the partitioning is fine. Fortunately, the $n \geq m$ panel size requirement may be circumvented with sparse data methods,

specifically ridge regression.

A recent theorem of Libgober (2021) provided the inspiration for our approach. If there are more worlds that experts $m > n$, a unique matrix W cannot be computed as $(X)^T X$ is not invertible. However, the matrix $(X)^T X - \varepsilon I$ is generically invertible for small ε, which leads to the well-known ridge regression approximation (used in Definition 6 below). Libgober assumed a possible world model with discrete signals, and proved that if the matrix X (in our notation) is known and if it includes among its rows all n possible beliefs, then the prior over $m > n$ worlds can be derived from eq. 15 in the limit as the regularization parameter ε goes to zero.

This limit result does not apply to our continuous belief model as all signals cannot be sampled. However, to the exent that the continuous distribution is well represented by a finite sample, the implication of Libgober's theorem may, for all practical purposes, survive, allowing the prior to be recovered with tolerable error. This suggests the following approximate mechanism.

Definition 6. Approximate Belief Decomposition mechanism. Let $\varepsilon = .000001$, or similar small value. Follow Steps 1-7 in Definition 4 with these modifications:

2. Combine reports into $n \times m$ matrices X, Y, and compute the *approximate local expectations matrix* $\tilde{W}$:

$$\tilde{W} = (X^T X - \varepsilon I)^{-1} X^T Y \tag{14}$$

Similarly, for each r combine reports excluding r into $(n-1) \times m$ matrices X^{-r}, Y^{-r}, and compute $\tilde{W}^{-r}$ as

$$\tilde{W}^{-r} = ((X^{-r})^T X^{-r} - \varepsilon I)^{-1} (X^{-r})^T Y^{-r} \tag{15}$$

5. Compute for all r the *approximate implied belief prediction* $\hat{y}^r \in \Delta(\Omega)$ such that:

$$\hat{y_i}^r \propto \sum_k x_k^r \tilde{w}_{ki}^{-r}, \;\; i = 1, \ldots, m \tag{16}$$

Although $\tilde{W}$ in Step 2 may not be a proper stochastic matrix, simulations indicate that the estimated prior (as in Definition 4, Step 4) tends to be close to the true prior, and only the prior is needed for inference. Figure II displays the accuracy of the estimated prior derived from simulated beliefs and belief predictions under randomly sampled possible world models for a few $n < m$ combinations (details in Supplemental Appendix). Accuracy is assessed by the KL-divergence between the actual

prior distribution $p(\omega)$ and the estimated prior distribution $\hat{p}(\omega)$. For each (n,m) pair, the figure indicates the KL-divergence distribution over 1000 trials, and shows the actual and estimated prior at the median accuracy level to enhance intuition for approximation quality. As one might expect, accuracy improves as the number of experts (n) increases and decreases with the number of possible worlds (m). Even with the minimal case of two experts, the deviation from true priors is small. The same simulations show that the error in the approximate implied predictions is similarly small. It appears reasonable to treat the error associated with these deviations as second-order relative to the background noise in human data.

V.B BD without conditional independence

Under conditional independence (Assumption 2), as panel size increases the BD posterior will converge to all probability mass on a single world state. The model setup assumes that truth is in principle knowable given sufficient respondents, since each respondent is guaranteed to provide additional information. The analyst should consider whether this is reasonable. If not, the question posed to the panel should be reworded to focus on available evidence rather than on ultimate states of the world. For example, the question 'True or False: Will the human race survive past 2100?', is not appropriate, because the true answer is not knowable irrespective of panel size. Instead, one should ask: 'Given current evidence, what is the likelihood of human survival past 2100?'

A sketch of the theoretical argument for this rewording is as follows. If Ω^* denotes the current state of evidence represented by discrete answers to the latter probabilistic question, and Ω the True/False ground truth in 2100, the variables are (roughly) arranged in a Markov chain, $\Omega \to \Omega^* \to S^r, S^{r'} \ldots$ We can then replace the strong assumption that $S^r, S^{r'} \ldots$ are conditionally independent on Ω with the weaker assumption that they are conditionally independent on Ω^*. Belief decomposition applied to Ω^* will deliver the full information posterior over Ω^*, and the mean of this distribution gives a single point forecast for Ω. The convergence of the posterior to a corner distribution, i.e, to complete confidence in one answer as more experts are sampled, would be unproblematic, as that answer would itself be a probability.

V.C BD when experts are not exchangeable

The 'identical' part of iid (Assumption 3, exchangeability) may fail if experts are not anonymous. Named individuals can bring reputations to the table, for particular expertise or disciplinary knowledge. To proceed with belief decomposition it is sufficient to select predictions about a single named

expert as the source for the common prior, and elicit the $n-1$ predictions of that expert's beliefs by others in the panel. (This is feasible only for $n > m$, hence the method does not apply to dyads). The matrix $\tilde{W}^{-r}$ and prior would be computed as before (equation 7 and 8), with X^{-r} and Y^{-r} given by the beliefs and predictions of these $n-1$ experts.

The matrix $\tilde{W}^{-r}$ also supports incentives for expert r, via equations 9 and 10. To add incentives for all players, one would need to elicit all pairwise predictions (as well as beliefs over world states), and compute for each expert r' their unique matrix $\tilde{W}^{-r'}$. A simpler version of this approach would be to identify experts by group characteristics only, and elicit predictions of beliefs across characteristics rather than names.

VI EMPIRICAL IMPLEMENTATION AND ASSESSMENT

While our objectives are primarily theoretical, we also provide an assessment of BD with published data. The datasets that we use involve science questions, with ex ante assigned difficulty, and questions about whether a named city is the state capital. These knowledge questions have several methodological advantages. In testing a new method, the answers must be verifiable, as these are. It is also desirable to have a spread in expertise, which is more likely for knowledge questions asked to a generic population than specialized questions posed to credentialed experts. The state capital datasets, in particular, contain many questions where majority opinion tends to be wrong, because a prominent city is not the capital. Examples include Philadelphia and Pennsylvania, or Chicago and Illinois. The presence of such items provides a sensitive test of whether a method can establish truth when most respondents hold an incorrect view.

We compare the performance of each method against classical and prediction based crowd wisdom algorithms for a range of sample sizes, evaluate individual-level BD scores as a measure of expertise, and evaluate the robustness of BD incentives to manipulation.

VI.A Methods tested

The common feature of all methods tested is that they do not require access to ground truth; however, there are differences in their input requirements. Some of the methods ask respondents to submit their beliefs as a discrete 'vote,' e.g., for the most likely world, and also to predict the distribution of votes in the panel. Assuming that experts can vote for one of N answers, we represent votes by the $0/1$ indicator vector $v^r = (v^r_1, \ldots, v^r_N) \in \Theta = \{(v_1, \ldots, v_N), v_i \in \{0,1\}, \sum_k v_k = 1\}$, and denote by

$z^r = (z_1^r, \ldots, z_N^r)$ the prediction of the distribution of votes excluding r.

Classical aggregation begins with simple majority vote, which does not deliver a probability but only the most likely world ω_{i°: $\omega_{i^\circ} = \arg\max_k \{\sum_r v_k^r\}$. The linear opinion pool (Stone, 1961) refines this by putting a mean probability, $\tilde{p}(\omega_i|x) = \frac{1}{n}\sum_{r=1}^{r=n} x_i^r$, on world ω_i (Makridakis and Winkler, 1983; Keuschnigg and Ganser, 2017). (With binary questions, this is equivalent to weighting answers by confidence). The logarithmic opinion pool (Cooke, 1991; P. Morris, 1977), puts the (normalized) geometric mean of respondent probabilities, $\tilde{p}(\omega_i|x) \propto (\Pi_{r=1}^{r=n} x_i^r)^{\frac{1}{n}}$, on world ω_i.

Prediction based methods start with the 'surprisingly popular' or SP algorithm (Prelec, Seung, and McCoy, 2017). In the binary case, SP selects the answer that maximizes the ratio of actual-to-predicted vote frequencies: $\omega_{i^\circ} = \arg\max_k \{\frac{\sum_r v_k^r}{\sum_r z_k^r}\}$.The 'surprisingly confident' or SC algorithm of Wilkening, Martinie, and Howe (2022) uses predictions about the beliefs (rather than votes) of others and selects ω_{i° according to: $\omega_{i^\circ} = \arg\max_k \{\frac{\sum_r x_k^r}{\sum_r y_k^r}\}$. This is a correction of the linear pool by the predicted linear pool. In contrast, BD computes the full information posterior: $\tilde{p}(\omega_i|x) \propto \frac{\Pi_{r=1}^{r=n} x_i^r}{\tilde{p}(\omega_i)^{n-1}}$.

The world prior can alternatively be estimated from predictions of the distribution of types or *Vote predictions* (VP) (Prelec and McCoy, 2022). VP requires the sample vote distribution $\bar{v} = (\bar{v}_1, \ldots, \bar{v}_N)$, $\bar{v}_i = \sum_r v_i^r/n$, and an estimate of the $N \times N$ average vote prediction matrix $\bar{Z} = [\bar{z}_{ij}]$, where $\bar{z}_{ij}$ is the average prediction of the proportion of respondents voting j by respondents themselves voting i. The prior over types $\tilde{p}(t^k)$ is estimated as the left unit eigenvector of $\bar{Z}$. The world prior is then estimated as $\tilde{p}(\omega_i) = \sum_k \bar{x}_{ik}\tilde{p}(t^k)$, where $\bar{x}_{ij}$ is the average probability put on world ω_j by respondents who vote for i. Since it requires that all types are sampled to compute the prior, VP only applies to dyads if the two respondents vote differently.

In the framework of this paper, all of these prediction based methods should identify the correct answer to a binary question in the large sample limit. For BD and VP this guarantee carries over to finite samples, in that the directional response will be consistent with the direction of the Bayesian full information posterior. SP and SC may give a directional response that is inconsistent with the full information posterior in the finite sample case.[3]

VI.B Datasets

Our analysis is based on three published datasets that, to our knowledge, are the only ones that elicit the four inputs needed to compare the tested methods: votes, probabilistic beliefs, predictions of

[3]For example, in the context of the Section II scenario with Experts A and B judging the probability of $\omega_1 = \textit{Success}$, if we obtain $(x_1^A, y_1^A) = (.8, .7)$, and $(x_1^B, y_1^B) = (.05, .10)$, then the SC ratio of $.85/.80 = 1.065$ would favor ω_1, but the computed full information posterior probability on ω_1 is only 33%.

average belief, and vote predictions.

The first dataset, 'WMH science' is from respondents (N=500) who answered subsets of 500 True/False science questions spanning grades 1 to 12 (Wilkening, Martinie, and Howe, 2022). The 500 problems were in five levels of difficulty (each corresponding to approximately two or three grade levels). Each level of difficulty contained fifty questions for which the correct answer is True, and fifty where the correct answer is False. Each respondent answered 20 questions from each level of difficulty, with question order randomized. Respondents were asked to (i) vote for True or False, (ii) predict the percentage of the sample voting True, (iii) estimate the probability that True is the correct answer, (iv) predict the average probability estimated by others.

The second dataset, 'WMH states' (Wilkening, Martinie, and Howe, 2022) asked respondents (N=100) whether a named city was the state capital for all 50 U.S. states. The named city was the largest city in the state. The same information as in WMH science was elicited from respondents.[4]

A third dataset, 'PSM states' similarly asked respondents (N=33) whether a named city was the state capital for all 50 U.S. states (Prelec, Seung, and McCoy, 2017). Respondents were asked to (i) vote for True or False, (ii) give their confidence from 50% to 100%, (iii) predict the percentage of the sample voting True, and (iv) predict the average confidence given by others. To impute belief predictions, we simply assume that a respondent predicts that everyone has the same confidence. That is, if c^r denotes the average confidence predicted by respondent r, then we impute $y^r = c^r z_1^r + (1-c^r)(1-z_1^r)$ where z_1^r is their prediction of the fraction of the sample voting True.

The first two studies were conducted online, and participants were not financially incentivized for accuracy to reduce motivation to look up the answer. The third study was conducted in the lab, without possibility of online search, and the top 20% of respondents with the most accurate Brier scores and the top 20% of respondents with the most accurate vote predictions earned a $25 bonus.

VI.C Aggregation results

VI.C.1 Aggregation results: full sample

Accuracy when all respondents are included provides an important initial benchmark across methods. Figure III displays the percent correct of different methods with the WMH science dataset,

[4]For the first dataset, Wilkening, Martinie, and Howe (2022) excluded 41 respondents who reported cheating or who did not complete the survey, and excluded the 11% of responses where a respondent's vote did not match their probabilistic answer. For the second dataset, a respondent's probabilistic judgment was forced to match a respondent's vote. For this dataset, eleven respondents who reported cheating or failed to complete the survey were excluded. We do the same.

where performance can be be assessed against ex ante problem difficulty. Percent correct is a valid measure because the 500 questions of this dataset are balanced in terms of the correct answer (as a whole, and at each difficulty level). For pairwise method comparisons, the difference between percent correct or Matthew's correlation coefficient is bootstrapped (5000 trials) across questions, and significance is evaluated with a z-test.

BD is evidently the most accurate method at every difficulty level, from 97% accuracy at level 1 to 84% accuracy at level 5. It is significantly more accurate than the surprisingly confident answer for the easiest difficulty level ($M = 13\%\,[7\%, 19\%]$, $z = 3.86$, $p < .001$). BD is significantly more accurate than the logarithmic pool for every difficulty level except for the easiest (2nd: $M = 9\%\,[2\%, 17\%]$, $z = 2.36$, $p = .02$; 3rd: $M = 14\%\,[6\%, 22\%]$, $z = 3.47$, $p < .001$; 4th: $M = 13\%\,[5\%, 22\%]$, $z = 2.94$, $p = .003$; 5th: $M = 17\%\,[7\%, 26\%]$, $z = 3.52$, $p < .001$), and significantly more accurate than VP for every difficulty level except for the two easiest (3rd: $M = 14\%\,[6\%, 22\%]$, $z = 3.35$, $p < .001$; 4th: $M = 13\%\,[4\%, 22\%]$, $z = 2.84$, $p = .005$; 5th: $M = 14\%\,[5\%, 23\%]$, $z = 3.21$, $p = .001$). Matched pair sign tests give similar results to evaluating bootstrapped percentage differences.

Unlike the science dataset, the two state datasets are unbalanced (17 questions for which True is correct, and 33 for which False is correct). To incorporate them into the comparison (Figure IV), requires an accuracy metric applicable to unbalanced data, such as Matthew's correlation coefficient (MCC). This is the metric used by Wilkening, Martinie, and Howe (2022). MCC ranges from -1 to +1, with higher score indicates higher accuracy.[5]

Pooling across the three datasets (600 questions total), BD remains the most accurate, and is approximately .09 MCC higher than the next most accurate method, the surprisingly confident answer ($M = .094\,[.049, .138]$, $z = 4.12$, $p < .001$). BD is significantly more accurate than vote predictions ($M = .145\,[.095, .193]$, $z = 5.64$, $p < .001$), and is. approximately .2 MCC higher than the logarithmic pool ($M = .195[.140, .246]$, $z = 7.22$, $p < .001$). The Supplemental Appendix provides these comparisons for each dataset separately.

As section VIII discusses, a failure of the information sharing assumptions can result in the BD full information posterior going to extremes. This is the case on the full samples in the datasets that we consider. For the 500 WMH science questions, BD gave a probability between .02 and .98 for only 11 of the questions, for WMH states on none of the questions, and for PSM states on 5 of the

[5]Matthew's correlation coefficient assumes binary categories, but the various aggregation methods occasionally result in ties. In such cases, we evaluate Matthew's correlation coefficient over double the number of questions where a tie is recorded as an answer once in each category, and non-ties occur twice in the doubled dataset.

questions. However, BD does have the most accurate Brier score for all datasets amongst methods that output a probability (details in Supplemental Appendix).

VI.C.2 Aggregation results: dyads

We evaluate BD on dyads constructed from arbitrary pairs of respondents making predictions about the sample average. For dyads, the 'confidence-based answer' is the answer given by the more confident respondent, and is also the answer selected by the linear or logarithmic opinion pool.

Section IV.B provides conditions that data elicited from a dyad satisfy if consistent with our model. Across all dyads and all questions, these conditions are satisfied by 33% of dyad-question pairs (of 1 648 352) for WMH science, 35% (of 195 800) for WMH states, and 50% (of 26 400) for PSM states. If a dyad fails these checks for a particular question, then BD defaults to the logarithmic pool and does no harm. BD seldom reverses the confidence-based answer on the dyads passing these checks: they give the same answer for 88% of these dyads for WMH science, 83% of these dyads for WMH states, and 90% of these dyads for PSM states. BD and the confidence-based answer thus have similar overall accuracy across all dyads (details in Supplemental Appendix).

To obtain a more refined picture of BD performance, we turn to the data subset where BD reverses the confidence-based answer and analyze this by dyad type (unanimous or split vote, favoring True or False). Because the distribution of correct True/False answers are now imbalanced, we also condition on the correct answer.. As Figure V shows, the percentage of correct BD reversals equals or exceeds the percentage of incorrect reversals for every dataset and dyad type. This difference is significant for all but two cases.

VI.C.3 Aggregation performance across sample sizes

Under BD, the impact of predictions on the full information posterior is filtered through the common prior estimate, $\tilde{p}(\omega)$, and this estimate should in theory be independent of sample size. To assess this, we examine performance at subsamples of different sized subsets, and compare against the linear pool and the logarithmic pool, which is directionally equivalent to BD with a uniform common prior. Except for PSM dyads and the full sample, the results are based on 5000 randomly sampled subsets (without replacement) as the number of possible subsets would be intractably large otherwise.

Figure VI displays how aggregation accuracy changes as a function of sample size. It is clear that the accuracy of traditional methods (the linear and logarithmic pools) quickly asymptotes, while the

accuracy of the BD posterior improves with sample size, without a clearcut asymptote. A reasonable conjecture is that the BD posterior benefits from a more stable estimate of the common prior. The Supplemental Appendix shows pairwise comparisons of methods at different sample sizes

VI.D Incentives

An effective incentive scheme should reward expertise, and also reward honesty apart from differences in expertise. These two desiderata are particularly important in the unverifiable domain, as experts should answer as if they would be scored against ground-truth even though ground truth will never be revealed. To make a preliminary assessment whether, apart from theory, this claim is reasonable in practice, we will evaluate whether respondents with greater ground truth accuracy indeed receive higher scores, and whether their scores would go down if their answer were systematically corrupted.

VI.D.1 Expertise incentives

For the analysis of incentives, we use the Kullback-Leibler divergence as the proper scoring rule, set $\alpha = .5$ and $\lambda = .5$, and compute the total score in the absence of side-payments. As Figure VII shows, BD respondent scores, and its components, capture expertise as measured by respondent Brier scores (correlations in figure). A respondent's total BD score, prediction accuracy, and implied prediction accuracy are significantly correlated with their Brier score for every dataset ($p < .001$, except $p = .001$ for the BD total score on WMH states). Scoring respondent beliefs against the BD full information posterior (by KL-divergence) also results in a high correlation ($p < .001$) with Brier scores, due to the accuracy of the BD full information posterior. The prediction garbling score is not significantly correlated with the Brier score (WMH science: $p = .76$, WMH states: $p = .07$, PSM states: $p = .07$). For WMH science and WMH states, the implied prediction accuracy score has a significantly higher correlation with respondent-level Brier scores than the prediction accuracy score, and the prediction accuracy score has a significantly higher correlation than the prediction garbling score (Steiger test for dependent correlations, details in Supplemental Appendix). For PSM states, implied prediction accuracy is significantly more diagnostic of respondent-level Brier scores than prediction garbling, but other differences are not significant. In the Supplemental Appendix, we show comparisons to other kinds of respondent scores.

Agreement with the full information posterior is the most sensitive correlate of ground-truth expertise. In theory, such agreement cannot provide belief incentives, as discussed in Section IV.

However, the deviation from incentive-compatibility diminishes with sample size, and disappears in the large sample limit.

VI.D.2 Honesty incentives

To evaluate the robustness of the honesty incentives provided by BD, we consider the effect of dishonest strategies on BD scores. Since there is no clear reference class of strategies that respondents could adopt in the hope of obtaining a better score, our approach is to consider a variety of 20 such strategies. Example strategies include symmetrically perturbing a respondent's belief by a fixed amount, and substituting vote predictions for beliefs. Consensus strategies put probability .70 (or .90) on the answer a respondent predicts is the majority answer, and similarly for contrarian strategies. Under mild versions of these strategies, respondents change their beliefs only if they predict their answer is shared with less than 30% of the sample for consensus strategies, and more than 70% of the sample for contrarian strategies. The beliefs or predictions of a single respondent are recoded under each strategy (with all other respondents' answers undisturbed), and a new score computed.

Table I shows the average score (across respondents) for a variety of strategies that manipulate beliefs, and a comparison at the individual respondent level between their score given their original beliefs and their score under the strategy (using a paired sample t-test). With Kullback-Leibler divergence and no side payments, scores are absolute-value penalties with zero the best score possible. An incentive system that rewards honesty would result in worse scores for the deceptive strategies compared to the original, presumably honest, beliefs. For WMH science, 16 of the strategies significantly worsen scores compared to the original beliefs (at the $p < .05$ level, Table I shows significance for each comparison), 3 show no significant difference, and 1 significantly improves respondent scores. For WMH states, 15 significantly worsen scores, 2 show no significant difference, and 3 significantly improve scores. For PSM states, 15 significantly worsen scores, 5 show no significant difference, and no strategy significantly improves scores. Only 3 strategies improve on reporting the original beliefs for any dataset, and no strategy improves scores for every dataset. (As expected, manipulation of predictions also reduces scores; these results are relegated to Supplemental Appendix).

The difference between the score of a strategy and the score of the original beliefs is largely driven by the mean absolute difference between the original beliefs and the manipulated beliefs (WMH science: $r = .97$, WMH states: $r = .92$, PSM states: $r = .90$; all $p < .0001$). For WMH science and PSM states, the mean absolute belief difference resulting from a strategy correlates with

the difference in scores, relative to baseline, of both implied prediction accuracy (WHM science: $r = .93$, PSM states: $r = .86$) and prediction garbling (WMH science: $r = .97$, PSM states: $r = .85$; all $p < .0001$). For WMH states, the change in scores, relative to baseline, as a function of the mean absolute belief difference is driven by prediction garbling ($r = .92$, $p < .0001$), rather than implied prediction accuracy ($r = .36$, $p = .11$).

VII OTHER RELATED WORK

VII.A Aggregation algorithms

The relevant comparison methods target unique questions, allow no communication or feedback, and have no access to ground truth. Extensive literatures deal with domains where one or more of these constraints are relaxed. For example, if respondents answer multiple questions, then one can weight them by their relative contribution to the group accuracy (Budescu and E. Chen, 2014) or their historical performance (Cooke and Goossens, 2008) or train a machine learning model, often a Bayesian hierarchical model, that aggregates judgments (Karabatsos and William H Batchelder, 2003; Oravecz, Anders, and William H Batchelder, 2013; Raykar et al., 2010; Merkle et al., 2016; Lee and Danileiko, 2014; Budescu and T. R. Johnson, 2011; Turner et al., 2014; McCoy and Prelec, 2023). If experts can answer sequentially after being given information about the opinions of others, one can use the extent to which an expert updates their opinion as a proxy for expertise (Zhang, 2020), or run a prediction market to aggregate opinions through the market price (Hanson, 2003; Arrow et al., 2008). Standard prediction markets require verifiable events, since participants are paid according to how the event being predicted resolves, but self-resolving prediction markets (Srinivasan, Karger, and Y. Chen, 2023) admit non-verifiable questions by terminating with some probability after each report, and paying respondents according to how much their report differs from a reference agent.

Traditionally, aggregation methods for single questions were based on respondents' beliefs, expressed either directly as subjective probabilities, or through votes and confidences (Cooke, 1991; Clemen and Winkler, 1999). With dyads, weighting votes by confidence yields a 'two heads are better than one' effect on consensually correct questions, but fails on questions where confidence tracks common knowledge, rather than the truth (Koriat, 2012). Large samples do not provide protection in 'wicked' environments (Hertwig, 2012), when extreme forecasts are given by people with little knowledge (Evgeniou et al., 2013) or misleading intuitions (Simmons et al., 2011).

Prediction based methods emerged to neutralize the shared-information bias. The 'surprisingly popular' or SP algorithm (Prelec, Seung, and McCoy, 2017), used predictions to define a handicap level for calibrating the strength of majority opinion. The SP theoretical results were asymptotic, and for non-binary questions, $m > 2$, required the assumption that respondents selecting the correct answer would assign more probability to that answer than other respondents. Wilkening, Martinie, and Howe (2022) proposed and tested the 'surprisingly confident' or SC algorithm for $m = 2$ that uses predictions about the beliefs (rather than votes) of others (see also Martinie, Wilkening, and Howe (2020)). The 'surprising overshoot' algorithm (Peker, 2023) proposed an estimator that uses the difference between how much an individual's own judgments overshoots the mean judgment and how much their predictions of others overshoots the mean judgment.

Like BD, the 'population-mean-based aggregation' (PMBA) of Y.-C. Chen, Mueller-Frank, and Pai (2021) elicits belief predictions and estimates the local expectations matrix W. However, instead of computing the prior as the stationary distribution of W, PMBA turns to asymptotic estimates of average beliefs. In the large sample limit, these estimates will converge to one column of the matrix W which then reveals the world state. Chen et al. show that the prior can be recovered from an infinite belief hierarchy, but also prove that any finite hierarchy leaves the prior indeterminate. Given this negative result, they acknowledge that the belief hierarchy approach seems difficult to implement for finite samples.

Palley and Soll (2019) and Palley and Satopää (2023) also address the shared information problem by tapping information contained in predictions about others. Their methods operate on continuous quantities, and assume that respondents receive a private and common signal which are combined linearly. The aim then is to weight respondents appropriately according to their inferred private signal. The pivoting method (Palley and Soll, 2019) pivots the aggregate answer away from the predictions of others thereby minimizing error under information structures with different combinations of informed and uninformed respondents. Palley and Satopää (2023) formalize the idea that respondents who are best at predicting others will themselves provide the most useful judgments, and show how to weight the point estimates of each respondent in a way that depends on a respondent's prediction of the mean estimate of the sample.

VII.B Incentives and mechanisms

There is a large literature on mechanisms for eliciting private information in finite player Bayesian games, and we review here only the strand that matches BD in its core assumptions, namely that

experts are disinterested (no preferences apart from scoring), and the mechanism is 'detail free' or 'robust' (Wilson, 1985) (no knowledge of the prior needed for implementation). Introducing non-scoring preferences would draw into discussion Bayesian mechanisms that exploit correlated signals (e.g., Myerson (1986), Crémer and McLean (1988), and Bergemann and S. Morris (2016)). Also outside the remit are mechanisms where the analyst knows the prior, (e.g., S. Johnson, Pratt, and Zeckhauser (1990) and Avery, Resnick, and Zeckhauser (1999)) and the peer-prediction method (Miller, Resnick, and Zeckhauser, 2005).

The Bayesian truth serum (BTS) (Prelec, 2004b), provided incentives for honest reporting of votes and vote predictions (but not beliefs) in large samples. Under BTS, all strict equilibria were either truth-telling or permutations of truth-telling (Cvitanic et al., 2020). Incentivizing votes and vote predictions has been the focus of subsequent work on finite 'truth serum' mechanisms. The Robust Bayesian truth serum (Witkowski and Parkes, 2012a) applied to binary question with $n \geq 3$, but required the insertion of a pairwise messaging step between vote submission and prediction. Radanovic and Faltings (2013) developed a method that accommodated non-binary multiple-choice questions under a 'self-prediction' condition, namely, that the highest prediction of the frequency of a given answer is expressed by people who endorse that answer. The divergence-based truth serum, in its 'non- parametric' version removed this limitation, by penalizing variance in predictions among respondents that submit the same vote (Radanovic and Faltings, 2014). This mechanism was the first to accommodate $n = 2$ dyads for incentivizing votes and vote predictions (Radanovic and Faltings (2014), Theorem 3). The choice-matching mechanism (Cvitanić et al., 2019) improved on Radanovic and Faltings (2014) in offering a decomposable score, but at the cost of requiring $n \geq m+2$, i.e., $n \geq 4$ for binary $m = 2$ questions. In an important simplification, the Bayesian market (Baillon, 2017) removed the need for vote predictions through a random-price market mechanism, but likewise required $n \geq 4$ and was limited to binary questions.

The main difference between this literature and our two mechanisms is that we formalize incentives for probabilistic beliefs, rather than type declarations, and explicitly model these beliefs over unknown world states. The one overlap in this respect is the population-mean-based aggregation (PMBA) mechanism of Y.-C. Chen, Mueller-Frank, and Pai (2021), where the incentives hold only in the large sample limit. Our mechanisms do require the common prior assumption, which several methods dealing with votes do not (e.g., Witkowski and Parkes (2012b) and Baillon (2017)).

Predictions about others have also been used in forecasting tournaments. In a tournament involving forecasts about the economy, politics, and covid-related events, (Himmelstein, Budescu,

and Ho, 2023) the accuracy of predictions about the forecasts made by others was found to pick out forecasters who also showed good performance on the actual forecasting task. Karger et al. (2021) suggest scoring forecasts against those of high-performing forecasters and showed that in a tournament involving short-term verifiable outcomes scoring forecasters in this way gave similar results to scoring forecasters against ground-truth and, in a tournament concerning hypothetical possible policies related to covid, resulted in a plausible and coherent ranking of policies.

VIII LIMITATIONS

A method that is valid in theory can fail in practice for any number of reasons: participants may exhibit psychological biases or have insufficient information, and more generally the theoretical assumptions of the model may not hold. The empirical results in Section VI are encouraging. However, some issues will require further research.

The directional improvement of BD over traditional methods (Figure VI) is a strong function of sample size. In theory, a two expert sample should identify the prior on binary questions. In practice, larger samples may be needed to stabilize estimates of the prior. We do not model respondent noise directly in this paper, but a related stream of literature explores this (McCoy and Prelec, 2023; Palley and Satopää, 2023; William H. Batchelder and Romney, 1988; Lee and Danileiko, 2014). For example, McCoy and Prelec (2023) develop a Bayesian hierarchical model that includes a respondent-level expertise parameter, and separate noise parameters for votes and vote predictions.

The directional accuracy of BD is excellent in comparison to other methods, indicating that the model assumptions hold sufficiently to improve inference. However, the posterior computed by BD is miscalibrated, albeit resulting in a better Brier score than other methods. The source of miscalibration almost surely arises from two assumptions basic to much of the literature: common knowledge and i.i.d. signals. These assumptions have several distinct issues.

First, information distributed among experts may not divide neatly into common knowledge and independent information. For example, information might be common knowledge among subsets of experts, but not among the entire sample. In principle, with small panels one could begin to address this by eliciting predictions of beliefs for each named expert separately, and computing pairwise common priors. Second, independent signals imply the existence of a very large pool of relevant information, as each expert is privy to a unique portion. Third, experts may not know whether their judgments are tapping independent or common sources. Concretely, a pair of grant

panel reviewers may not know how much common information they share, that is, whether they are from the same sub-discipline, on top of the same literature, and so forth. Furthermore, the joint distribution of signals and world states is unlikely to be known to all experts independent of the information that experts possess. Higher-order uncertainty about the evidential base, i.e., source uncertainty, could be resolved by giving experts a mechanism for labeling their evidence, and sharing the labels. However, if communication is not feasible or not desirable, one could lessen the source uncertainty problem by expanding the original question into a query about evidence available to the expert sample, as suggested in V.B.

With respect to incentives, there is a potential risk insofar as the Belief Decomposition mechanism does not link probabilities to objective frequencies. Because belief predictions are scored against beliefs, experts could distort both scales consistently, for example, by treating a probability of 80% as 90% for both beliefs and belief predictions. An alternative mechanism that used vote predictions could score predictions against the empirical vote distribution, which cannot be distorted without cost (McCoy and Prelec, in prep.).

IX CONCLUSION

We have presented collective inference as the task of outputting the full information posterior given a set of reports that are governed by incentives, and have defined a mechanism that enables this. The Belief Decomposition mechanism elicits predictions about the beliefs of others, and uses these predictions to incentivize respondent beliefs even when ground truth is non-verifiable. Predictions about the beliefs of others are also elicited because beliefs alone, however correctly expressed, are not sufficient for inference. What requires aggregation is the mix of private and shared information possessed by respondents, not individual beliefs. When expert probabilities rehearse the same evidence, an apparent consensus may be directionally misleading, as illustrated by the two expert example in Section II. The observation that simple averaging overweights shared information is not novel, but is largely ignored in practice. Classical methods such as averaging expert probabilities or weighting them by confidence remain the default, because of familiarity and ease of use. To have impact on practice, any alternative should not only be guaranteed in theory, but also simple to implement with respect to elicitation and computation.

It is thus advantageous that the computation involved in BD is fairly straightforward. A local expectations matrix is computed by linear regression, which yields the common prior through its

stationary distribution. This common prior is then incorporated into the logarithmic pool. Non-verifiable subjective probabilities can be scored from implied predictions obtained from this matrix. An approximate BD mechanism can be implemented when there are fewer experts than world states. The empirical results are promising on the data available so far: Belief decomposition provides substantial accuracy improvements in many cases, and otherwise does no harm.

The simplicity of BD expands the range of situations where it might be applied. From an expert's perspective, they simply supply their belief and belief prediction, and receive a score when the whole group has done so. In principle, BD could be used whenever expert opinion from small or large groups is needed (Surowiecki, 2005; Sunstein, 2006). Today, many collective judgment situations do not involve the formal elicitation and aggregation of beliefs, but this is arguably a missed opportunity. For example, one could routinely elicit, in a light-touch manner, judgments from a patient's medical team about a diagnosis or best treatment, and then aggregate and score these judgments. Such a procedure would enhance later group discussion by surfacing common knowledge, and suggesting, as a target for debate, a group answer, that may not align with any individual's beliefs. Similarly, referees of journal articles or grant proposals could be asked to provide predictions as well as personal judgments, along the lines of our example in Section II.

Forecasting is a natural application domain, especially for non-verifiable forecasts of the remote future, or the effect of alternative policy choices (Morgan, 2014). In geopolitical forecasting, in particular, human judgment will continue to play a critical role (Tetlock and Gardner, 2016; Mellers et al., 2023). In all these settings, the respondent-level scores provided by BD could be used by internal teams making forecasts on a regular basis. Scores need not be tied to compensation, but would provide reputational incentives (free of consensus bias), and track expertise over time.

One might ask why incentives are needed if experts are not motivated to misreport probabilities. One reply is that this represents a detail-free modeling strategy that starts with necessary rather than sufficient incentive conditions. Mechanisms that do not support a strict equilibrium even if players are merely indifferent about honesty, certainly cannot be trusted in situations when players might have a strong motive to deceive. Whether these incentives are also sufficient will depend on details such the strength of the deception motive.

A more substantive reply is that the primary objective of incentives here is to reward identification and communication of information, in a setting where the terms in which events are described are themselves sometimes open to debate. The enemy is vagueness and confusion rather than deliberate deception. In this respect, a collective inference mechanism is similar to proper scoring rules,

which likewise do not explicitly model the utility of deception. Both are instruments for rewarding meta-cognition, the ability to discriminate, interpret and communicate cognitive states (Fleming and Frith, 2014; Heyes et al., 2020). The relevant cognitive states here are states of belief, but with a further complication. The individual has to consider the sources of belief, and separate information that gives a unique insight from information that is likely shared. This may be an unfamiliar 'signal-extraction' task; the mechanism score is designed to help learn how to do it well.

There is of course an existing (and well promoted) institution that also rewards individual skill in distinguishing own from shared information. In markets, respondents trade implicitly on their beliefs about the difference between their own judgment and that of others. In contrast, BD has respondents explicitly report both these judgments. Standard prediction markets can be used for verifiable events, but respondents are only paid when the event resolves.

Biases such as overconfidence and base-rate neglect in individual judgment are well documented, and methods exist to attempt to correct such biases. Incentives can help by encouraging deliberation and enabling feedback even without verifiable ground-truth. But, for expert judgments to be maximally useful, this is not sufficient. Effective decision making requires the implied best collective probability, and many methods used today do not provide this even in the absence of individual bias. The development of collective inference mechanisms, including methods of belief decomposition, should encourage eliciting expert judgment in more situations, and enable more successful use of such judgments.

Appendix I Proofs of Propositions 1 and 2

Proof of Proposition 1.

We first derive the world prior with a pair of Lemmas.

Lemma 2. *Under Assumptions 1-4, the local expectations matrix W can be computed from a generic random sample of $n \geq m$ honest beliefs $X = [x^1, \ldots, x^n]$ and corresponding honest predictions $Y = [y^1, \ldots, y^n]$ with $W = \tilde{W}$.*

Proof. The local expectation matrix W is defined as

$$W = [w_{ij}] = [\int_{s^r} p(\Omega = \omega_j | S^r = s^r) p(S^r = s^r | \Omega = \omega_i)]$$

since Assumption 3 (exchangeability) implies that such local expectations do not require the choice of a particular expert.

We can express Y in terms of X and W. Letting t denote any expert $t \neq r$,

$$\begin{aligned} y^r_j &= \int_{s^t} p(\Omega = \omega_j | S^t = s^t) \sum_{\omega_k} p(S^t = s^t | \Omega = \omega_k) p(\Omega = \omega_k | S^r = s^r) \\ &= \sum_{\omega_k} p(\Omega = \omega_k | S^r = s^r) \int_{s^t} p(\Omega = \omega_j | S^t = s^t) p(S^t = s^t | \Omega = \omega_k) \\ &= \sum_k x^r_k w_{kj} \end{aligned}$$

Belief predictions are therefore a garbling of beliefs via W:

$$Y = XW$$

Since the columns of a generic X are linearly independent (by Assumption 4, full support) and $n \geq m$, one can compute t $(X^T X)^{-1} X^T$, and compute $\tilde{W}$ as:

$$\tilde{W} = (X^T X)^{-1} X^T Y$$

with $\tilde{W} = W$. Since W is row stochastic so too is $\tilde{W}$. ■

Lemma 3. *Under Assumptions 1-4, the world prior, $p(\Omega = \omega_j),\ j = 1,\ldots,m$, can be computed from a generic random sample of $n \geq m$ honest beliefs $X = [x^1,\ldots,x^n]$ and corresponding honest predictions $Y = [y^1,\ldots,y^n]$. Specifically $p(\omega) = \tilde{p}(\omega)$ where $\tilde{p}(\omega)$ is the unique left eigenvector of $\tilde{W}$.*

Proof. With $W = \tilde{W}$ computed from X, Y (Lemma 2), we have:

$$\begin{aligned}
p(\Omega = \omega_j) &= \int_{s^t} p(\Omega = \omega_j | S^t = s^t) p(S^t = s^t) \\
&= \int_{s^t} p(\Omega = \omega_j | S^t = s^t) \sum_k p(S^t = s^t, \Omega = \omega_k) \\
&= \sum_k \int_{s^t} p(\Omega = \omega_j | S^t = s^t) p(S^t = s^t, \Omega = \omega_k) \\
&= \sum_k [\int_{s^t} p(\Omega = \omega_j | S^t = s^t) p(S^t = s^t | \Omega = \omega_k)] p(\Omega = \omega_k) \\
&= \sum_k w_{kj} p(\Omega = \omega_k)
\end{aligned}$$

The prior $p(\omega)$ is therefore the stationary distribution of the irreducible Markov chain defined by the row stochastic matrix $W = \tilde{W}$, i.e., it is the unique left eigenvector of $\tilde{W}$, and $p(\omega) = \tilde{p}(\omega)$. ■

Lemma 4. *If μ is a Bayesian Nash equilibrium with signal-pooling strategies, $\mu^r(s) = (x^r, y^r)$ for some constant probability vectors x^r, y^r, then $\tilde{W}$ cannot be computed or $\tilde{W}$ is either the identity matrix or is not a stochastic matrix.*

Proof. Because x^r are all common knowledge when signal-pooling, predictions are perfect, $y^r = E[\bar{x}^{-r} | s^r] = \bar{x}^{-r}$. If x^r are identical for all r, then $Y = X$, and $\tilde{W} = (X^T X)^{-1} X^T Y$ is the identity matrix. If sufficient x^r are identical that $X^T X$ is nearly singular, then the computation of its inverse is unstable and we say that $\tilde{W}$ cannot be computed. Suppose that this is not the case. If beliefs are not identical for a pair of experts r, t, the difference in their predictions for world i has sign opposite to the difference in their beliefs:

$$y_i^r - y_i^t = \bar{x}_i^{-r} - \bar{x}_i^{-t} = \frac{1}{n-1}(x_i^t - x_i^r) \tag{17}$$

We may also express prediction differences in terms of garbled beliefs, using matrix $\tilde{W}$:

$$\begin{aligned} y_i^r - y_i^t &= \sum_k \tilde{w}_{ki} x_k^r - \sum_k \tilde{w}_{ki} x_k^t \\ &= \sum_k \tilde{w}_{ki} (x_k^r - x_k^t) \\ &= -\sum_k (n-1) \tilde{w}_{ki} (y_k^r - y_k^t) \end{aligned} \tag{18}$$

where the last line is obtained by substituting from eq. 17. The matrix $(n-1)\tilde{W}$ defines a vector reflection through the origin. Consequently, $\tilde{W}$ itself will contain some negative coefficients, i.e., it will not be a stochastic matrix. ■

Proposition. *If $\mathscr{P}$ satisfies Assumptions 1-5, the number of experts exceeds the number of worlds ($n > m$), and $\mathscr{M} = \langle \mathscr{A}, (V^r)_{r=1}^n, f \rangle$ is constructed according to Definition 4, then $\mathscr{M}$ is a pooling-proof collective inference mechanism on $\mathscr{P}$.*

Proof. For any possible world model $p \in \mathscr{P}$, define the honest strategy profile $\sigma = (\sigma^1, \ldots, \sigma^n)$, $\sigma^r : S^r \to \mathscr{A} = \Delta(\Omega) \times \Delta(\Omega)$, as:

$$\begin{aligned} \sigma^r(s^r) &= (x^r(s^r), y^r(s^r)) \\ x^r(s^r) &= p(\omega | s^r) \\ y^r(s^r) &= \int_{s^t \neq s^r} p(\omega | s^t) \sum_{\omega'} p(s^t | \omega') p(\omega' | s^r)) \end{aligned}$$

and let σ^{-r} denote the honest strategy profile excluding r.

We prove that σ is a strict Bayesian equilibrium. We first consider each term of the scoring function V when $C = 1$, which is the case for σ. The score involves four terms, repeated here for convenience:

$$\begin{aligned} &(1)\ \textit{prediction score} && \alpha u(\bar{x}^{-r}, y^r) \\ &(2)\ \textit{imputed prediction score} && (1-\alpha) u(\bar{x}^{-r}, \hat{y}^r) \\ &(3)\ \textit{prediction garbling score} && \lambda u(y^r, \hat{y}^r) \\ &(4)\ \textit{side payment} && -u(\bar{x}^{-r}, \tilde{p}(\omega)^{-r}) \end{aligned} \tag{19}$$

The side-payment may be set aside as as it does not involve $\sigma(s^r)$. The prediction score is strictly maximized by $y^r = E_{\sigma^{-r}}[\bar{x}^{-r} | s^r]$ because u is a proper scoring rule. Since experts $t \neq r$ are in honest

equilibrium, $x^t = p(\omega|s^t)$, and $\bar{x}^{-r} = (n-1)^{-1}\sum_{t\neq r}\sum_{s^t} p(\omega|s^t)$ giving:

$$\begin{aligned} E_{\sigma^{-r}}[\bar{x}^{-r}|s^r] =& (n-1)^{-1}\sum_{t\neq r}\int_{s^t} p(\omega|s^t)\sum_k p(s^t|\omega_k)p(\omega_k|s^r) \\ =& \int_{s^t} p(\omega|s^t)\sum_k p(s^t|\omega_k)p(\omega_k|s^r) \end{aligned}$$

as $s^t = s^{t'} \implies p(\omega|s^t) = p(\omega|s^{t'})$, by Assumption 3 (exchangeability). Therefore honest belief predictions $y^r(s^r) = \int_{s^t} p(\omega|s^t)\sum_k p(s^t|\omega_k)p(\omega_k|s^r)$ strictly maximize the first term of V.

Turning to the second and third terms, observe that if all experts $t \neq r$ are in honest equilibrium, then matrices X^{-r} and Y^{-r} will respectively contain honest belief and prediction vectors. Therefore, $\tilde{W}^{-r}$ will compute the correct local expectations, i.e. $\tilde{W}^{-r} = W$, (noting that $n-1 \geq m$):

$$\tilde{w}_{ij}^{-r} = \sum_k p(\omega_j|s_k)p(s_k|\omega_i)$$

Reporting beliefs honestly results in a match between imputed predictions and stated predictions:

$$\begin{aligned} x_i^r = p(\omega_i|s^r) \implies \hat{y_j}^r(x^r) =& \sum_i x_i^r \tilde{w}_{ij}^{-r} \\ =& \sum_i p(\omega_i|s^r)\sum_k p(\omega_j|s_k)p(s_k|\omega_i) \\ =& \sum_k p(\omega_j|s_k)\sum_i p(s_k|\omega_i)p(\omega_i|s^r) \\ =& y_j^r(s^r) \end{aligned}$$

Honest beliefs simultaneously maximize the imputed prediction score and prediction garbling score. ($u(y^r,\hat{y}^r)$, also involves y^r, but that is irrelevant as $u(y^r,\hat{y}^r) = 0$ attained at $y^r = \hat{y}^r$ is the best possible score over all pairs $(y^r,\hat{y}^r)$). By Assumption 5, stochastic relevance, only honest beliefs would yield the match, $\hat{y}^r(x^r) = y^r(s^r)$, which proves that honest beliefs strictly maximize expected score.

If $C = 0$ under the deviation of r from σ we show that this results in a lower score than σ. Because u is a strictly proper score, and $\hat{y}^r = y^r$, $\tilde{p}^{-r}(\omega) = p(\omega)$ in equilibrium, it follows that $E[u(\bar{x}^{-r},y^r)|s^r] = E[u(\bar{x}^{-r},\hat{y}^r)|s^r] \geq E[u(\bar{x}^{-r},p(\omega))|s^r]$, with inequality strict unless $E[\bar{x}^{-r}|s^r] = p(\omega)$. Expected scores are therefore strictly positive in equilibrium for all experts except those whose predictions match the prior (by Assumption 4 full support, these experts have measure zero, but that is not critical for the argument here). However, $C = 0$ would yield a negative score as u is a proper penalty. Hence, σ is a strict Bayesian equilibrium.

We now show that σ is pooling-proof. Let $\mu(s) = (a^1, \ldots, a^n)$ with $\mu^r(s) = (x^r, y^r)$ for some constant probability vectors x^r, y^r, so that μ is a Bayesian Nash equilibrium with signal-pooling. By Lemma 4, if $\tilde{W}$ is computable then the matrix $\tilde{W}$ is either the identity matrix or is not row stochastic, which would set $C = 0$, and yield a negative score. Because expected scores are non-negative under σ, pooling would yield strictly less than the honest equilibrium.

We conclude by showing that the inference function applied to $\sigma(s)$ computes the full information posterior. By Lemma 2, under a generic truth-telling equilibrium, $\tilde{W}$ is a computable row-stochastic matrix, and since $n-1 \geq m$ so too is $\tilde{W}^{-r}$ for every r. Hence, $C = 1$. By Bayes' rule and conditional independence (Assumption 2), the full information posterior $p(\omega_i|s)$ is proportional to:

$$\begin{aligned} p(\omega_i|s) &\propto p(s|\omega_i)p(\omega_j \\ &= \prod_{r=1}^{n} p(s^r|\omega_i)p(\omega_i) \\ &= \prod_{r=1}^{n} \frac{p(\omega_i|s^r)p(s^r)}{p(\omega_i)} p(\omega_i) \\ &\propto \prod_{r=1}^{n} x_i^r \tilde{p}(\omega_i)^{n-1} \end{aligned}$$

where the last line follows from Lemma 3, which proves that $\tilde{p}(\omega) = p(\omega_i)$, and the definition of σ_p. Normalization yields the full information posterior as in Definition 4(7). ■

Proof of Proposition 2.

Proposition. *If $\mathscr{P}$ satisfies Assumptions 1-5, the number of experts is greater or equal to the number of worlds ($n \geq m$), and $\mathscr{M} = \langle \mathscr{A}, (V^r)_{r=1}^n, f \rangle$ is constructed according to Definition 5, then $\mathscr{M}$ is a pooling-proof collective inference mechanism on $\mathscr{P}$.*

Proof. We define honest strategies and profiles $\sigma, \sigma^r, \sigma^{-r}$ as in Proposition 1. Assuming that $t \neq r$ are in honest equilibrium, then honest predictions $y^r(s^r)$ strictly maximizes the expectation $E[V^r(x,y)|s^r]$ with respect to y^r.

In a generic honest equilibrium, $C = 1$ as $\tilde{W}$ will be a computable row stochastic matrix and different from the identity matrix. To prove that the equilibrium is strict with respect to beliefs, we show that any deviation from $x^r(s^r) = p(\omega|s^r)$ creates some risk of non-computability, $C = 0$. In the binary case, the deviation risks misalignment of beliefs and predictions, as discussed in the Section II example. The proof generalizes this to $m > 2$.

Because honest predictions are a garbling of honest beliefs by the stochastic matrix W, given any two experts, r and t, the quadratic distance between their honest beliefs should be greater than the distance between their honest predictions:

$$\sum_j (y_j^r - y_j^t)^2 = \sum_j \sum_k (w_{kj}(x_j^r - x_j^t))^2 < \sum_j (x_j^r - x_j^t)^2$$

By contraposition, if matrices X and Y contain vector pairs (x^r, y^r) and (x^t, y^t) for which above inequality fails, then $\tilde{W}$ will not compute as a row stochastic matrix. Suppose now that r reports the dishonest belief $x^r \neq p(\omega|s^r)$, and assume that $\tilde{W}$ is a computable row stochastic matrix based on dishonest beliefs by r but with all other inputs honest. We show that the inequality will fail if an honest expert t has beliefs that fall in the neighborhood of the dishonest beliefs reported by r, that is, if $x^r \approx p(\omega|s^t) = x^t$ but $y^r \neq y^t$. Consider the inputs of an honest expert t whose beliefs exactly match r's stated beliefs, $x^t = x^r$. In that case,

$$y_j^t = \sum_k w_{kj} p(\omega_k|s^t) = \sum_k w_{kj} x_k^t$$
$$y_j^r = \sum_k w_{kj} p(\omega_k|s^r) \neq \sum_k w_{kj} x_k^r$$

where the inequality in the second line reflects $x_k^r \neq p(\omega_k|s^r)$ and the fact that W has full rank. The quadratic distance between x^r and x^t is zero, while the quadratic distance between y^r and y^t is positive, $\sum_j (\sum_k w_{kj}(p(\omega_k|s^t) - p(\omega_k|s^r))^2 = K > 0$. By continuity, as $x^t \to x^r$, $\sum_k (x_k^r - x_k^t)^2 \to 0$, but $\sum_j (y_j^r - y_j^t)^2 \to K > 0$, i.e., for all x^t in the neighborhood of x^r we will have $\sum_j (y_j^r - y_j^t)^2 > \sum_k (x_k^r - x_k^t)^2$, and the estimated $\tilde{W}$ will not compute as a stochastic matrix. By full support, if x^r is not honest, there is a positive probability that x^t will fall in the neighborhood of beliefs reported by r, and therefore a positive probability of the event $C = 0$.

Pooling-proof is secured by Lemma 4 as in proof of Proposition 1. The proof that the inference function computes the full information posterior is the same as in Proposition 1. ■

References

Arieli, Itai, Yakov Babichenko, and Rann Smorodinsky (2020). "Identifiable information structures". *Games and Economic Behavior* 120, pp. 16–27.

Arrow, Kenneth J et al. (2008). *The promise of prediction markets*.

Avery, Christopher, Paul Resnick, and Richard Zeckhauser (1999). "The market for evaluations". *American Economic Review* 89.3, pp. 564–584.

Baillon, Aurélien (2017). "Bayesian markets to elicit private information". *Proceedings of the National Academy of Sciences* 114.30, pp. 7958–7962.

Batchelder, William H. and A.K. Romney (1988). "Test theory without an answer key". *Psychometrika* 53.1, pp. 71–92.

Bergemann, Dirk and Stephen Morris (2016). "Bayes correlated equilibrium and the comparison of information structures in games". *Theoretical Economics* 11.2, pp. 487–522.

Budescu, David V and Eva Chen (2014). "Identifying expertise to extract the wisdom of crowds". *Management Science*.

Budescu, David V and Timothy R Johnson (2011). "A model-based approach for the analysis of the calibration of probability judgments." *Judgment & Decision Making* 6.8.

Chen, Yi-Chun, Manuel Mueller-Frank, and Mallesh M Pai (2021). "The wisdom of the crowd and higher-order beliefs". *arXiv preprint arXiv:2102.02666*.

Clemen, Robert T and Robert L Winkler (1999). "Combining probability distributions from experts in risk analysis". *Risk analysis* 19.2, pp. 187–203.

Condorcet, Marquis de (1785). *Essay sur l'application de l'analyse de la probabilité des decisions: redues et pluralité des voix*. l'Imprimerie Royale.

Cooke, Roger (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, USA.

Cooke, Roger and Louis LHJ Goossens (2008). "TU Delft expert judgment data base". *Reliability Engineering & System Safety* 93.5, pp. 657–674.

Crémer, Jacques and Richard P McLean (1988). "Full extraction of the surplus in Bayesian and dominant strategy auctions". *Econometrica: Journal of the Econometric Society*, pp. 1247–1257.

Cvitanic, Jaksa et al. (2020). "Incentive-Compatible Surveys via Posterior Probabilities". *Theory of Probability & Its Applications* 65.2, pp. 292–321.

Cvitanić, Jakša et al. (2019). "Honesty via choice-matching". *American Economic Review: Insights* 1.2, pp. 179–192.

Evgeniou, Theodoros et al. (2013). "Competitive dynamics in forecasting: The interaction of skill and uncertainty". *Journal of Behavioral Decision Making* 26.4, pp. 375–384.

Fleming, Stephen M and Christopher D Frith (2014). *The cognitive neuroscience of metacognition*. Springer.

Galton, F. (1907). "Vox populi". *Nature* 75, pp. 450–451.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American statistical Association*, pp. 359–378.

Grace, Katja et al. (2024). "Thousands of AI authors on the future of AI". *arXiv preprint arXiv:2401.02843*.

Hanson, Robin (2003). "Combinatorial information market design". *Information Systems Frontiers* 5, pp. 107–119.

Hertwig, Ralph (2012). "Tapping into the Wisdom of the Crowd-with Confidence". *Science* 336.6079, pp. 303–304.

Heyes, Cecilia et al. (2020). "Knowing ourselves together: the cultural origins of metacognition". *Trends in cognitive sciences* 24.5, pp. 349–362.

Himmelstein, Mark, David V Budescu, and Emily H Ho (2023). "The wisdom of many in few: Finding individuals who are as wise as the crowd." *Journal of Experimental Psychology: General*.

Johnson, Scott, John W Pratt, and Richard Zeckhauser (1990). "Efficiency despite mutually payoff-relevant private information: The finite case". *Econometrica: Journal of the Econometric Society*, pp. 873–900.

Karabatsos, George and William H Batchelder (2003). "Markov chain estimation for test theory without an answer key". *Psychometrika* 68.3, pp. 373–389.

Karger, Ezra et al. (2021). "Reciprocal scoring: A method for forecasting unanswerable questions". *Available at SSRN 3954498*.

Keuschnigg, Marc and Christian Ganser (2017). "Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule". *Management science* 63.3, pp. 818–828.

Koriat, A. (2012). "When Are Two Heads Better than One and Why?" *Science* 336.6079, pp. 360–362.

Lee, Michael D and Irina Danileiko (2014). "Using cognitive models to combine probability estimates". *Judgment and Decision Making* 9.3, pp. 259–273.

Libgober, Jonathan (2021). "Hypothetical Beliefs Identify Information". *arXiv:2105.07097*.

Makridakis, Spyros and Robert L Winkler (1983). "Averages of forecasts: Some empirical results". *Management Science* 29.9, pp. 987–996.

Martinie, Marcellin, Tom Wilkening, and Piers DL Howe (2020). "Using meta-predictions to identify experts in the crowd when past performance is unknown". *Plos one* 15.4, e0232058.

McCoy, John and Drazen Prelec (in prep.). "Vote prediction mechanisms".

— (2023). "A Bayesian Hierarchical Model of Crowd Wisdom Based on Predicting Opinions of Others". *Management Science*.

Mellers, Barbara A et al. (2023). "Human and Algorithmic Predictions in Geopolitical Forecasting: Quantifying Uncertainty in Hard-to-Quantify Domains". *Perspectives on Psychological Science*, p. 17456916231185339.

Merkle, Edgar C et al. (2016). "Item response models of probability judgments: Application to a geopolitical forecasting tournament." *Decision* 3.1, p. 1.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser (2005). "Eliciting informative feedback: The peer-prediction method". *Management Science* 51.9, pp. 1359–1373.

Morgan, M Granger (2014). "Use (and abuse) of expert elicitation in support of decision making for public policy". *Proceedings of the National academy of Sciences* 111.20, pp. 7176–7184.

Morris, P.A. (1977). "Combining expert judgments: A Bayesian approach". *Management Science* 23.7, pp. 679–693.

Myerson, Roger B (1986). "Multistage games with communication". *Econometrica: Journal of the Econometric Society*, pp. 323–358.

Oravecz, Zita, Royce Anders, and William H Batchelder (2013). "Hierarchical bayesian modeling for test theory without an answer key". *Psychometrika*, pp. 1–24.

Palley, Asa B and Ville A Satopää (2023). "Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions". *Management Science*.

Palley, Asa B and Jack B Soll (2019). "Extracting the wisdom of crowds when information is shared". *Management Science* 65.5, pp. 2291–2309.

Peker, Cem (2023). "Extracting the collective wisdom in probabilistic judgments". *Theory and Decision* 94.3, pp. 467–501.

Prelec, Drazen (Oct. 2004a). “A Bayesian truth serum for subjective data”. *Science* 306.5695, pp. 462–6. DOI: 10.1126/science.1102081.

— (2004b). “A Bayesian truth serum for subjective data”. *Science* 306.5695, pp. 462–466.

Prelec, Drazen and John McCoy (2022). “General identifiability of possible world models for crowd wisdom”.

Prelec, Drazen, H Sebastian Seung, and John McCoy (2017). “A solution to the single-question crowd wisdom problem”. *Nature* 541.7638, pp. 532–535.

Radanovic, Goran and Boi Faltings (2013). “A robust bayesian truth serum for non-binary signals”. : *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI" 13)*. EPFL-CONF-197486, pp. 833–839.

— (2014). “Incentives for truthful information elicitation of continuous signals”. : *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14)*. CONF, pp. 770–776.

Raykar, Vikas C et al. (2010). “Learning from crowds”. *Journal of Machine Learning Research* 11.Apr, pp. 1297–1322.

Simmons, Joseph P et al. (2011). “Intuitive biases in choice versus estimation: implications for the wisdom of crowds”. *Journal of Consumer Research* 38.1, pp. 1–15.

Srinivasan, Siddarth, Ezra Karger, and Yiling Chen (2023). “Self-Resolving Prediction Markets for Unverifiable Outcomes”. *arXiv preprint arXiv:2306.04305*.

Stone, Mervyn (1961). “The opinion pool”. *The Annals of Mathematical Statistics*, pp. 1339–1342.

Sunstein, C.R. (2006). *Infotopia: How many minds produce knowledge*. Oxford University Press, USA.

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

Tetlock, Philip E and Dan Gardner (2016). *Superforecasting: The art and science of prediction*. Random House.

Turner, Brandon M et al. (2014). “Forecast aggregation via recalibration”. *Machine Learning* 95.3, pp. 261–289.

Wilkening, Tom, Marcellin Martinie, and Piers DL Howe (2022). “Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems”. *Management Science* 68.1, pp. 487–508.

Wilson, Robert (1985). *Game-theoretic analyses of trading processes*. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Witkowski, Jens and David Parkes (2012a). “A robust bayesian truth serum for small populations”. : *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1.

— (2012b). “Peer prediction without a common prior”. : *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, pp. 964–981.

Zhang, Yunhao (2020). “The Revealed Expertise Algorithm: Leveraging Advice-taking to Identify Experts and Improve Wisdom of Crowds”. *Available at SSRN 3739192*.

Tables

Strategy	WMH science	WMH states	PSM states	Strategy	WMH science	WMH states	PSM states
Original answers	**.208**	**.213**	**.102**	Perturb by .05	.209**	.212	.102
Perturb by .10	.215***	.214*	.102	Perturb by .15	.224***	.217***	.103**
Perturb by .20	.237***	.222***	.104***	Perturb by .25	.252***	.228***	.106***
Always p = .90	.522***	.302***	.130***	Always p = .10	.379***	.306***	.140***
Reverse probabilities	.632***	.353***	.165***	Randomly sample	.361***	.262***	.124***
Consensus p = .70	.206	.202***	.106*	Consensus p = .90	.233***	.224**	.109***
Mild consensus p = .70	.210	.204***	.104	Mild consensus p = .90	.207	.208	.101
Contrarian p = .70	.396***	.278***	.128***	Contrarian p = .90	.650***	.375***	.154***
Mild contrarian p = .70	.374***	.270***	.121***	Mild contrarian p = .90	.568***	.340***	.136***
Vote predictions	.569***	.323***	.144***	Belief predictions	.577***	.333***	.133***
Reverse belief predictions	.187***	.193***	.104				

Table I: Results of strategies that manipulate the beliefs of each respondent in turn. Higher scores are worse. Each cell shows the average score for a strategy (across respondents and questions) on a dataset, and a comparison (using a paired sample t-test) of respondent scores under the strategy compared to honestly reporting elicited beliefs. We denote $p < .05$ with *, $p < .01$ with **, and $p < .001$ with ***. For the majority of cases, a deceitful strategy worsens scores. We highlight in red the few cases where a strategy significantly improves scores compared to no manipulation. No strategy significantly improves scores for all three datasets.

Figures

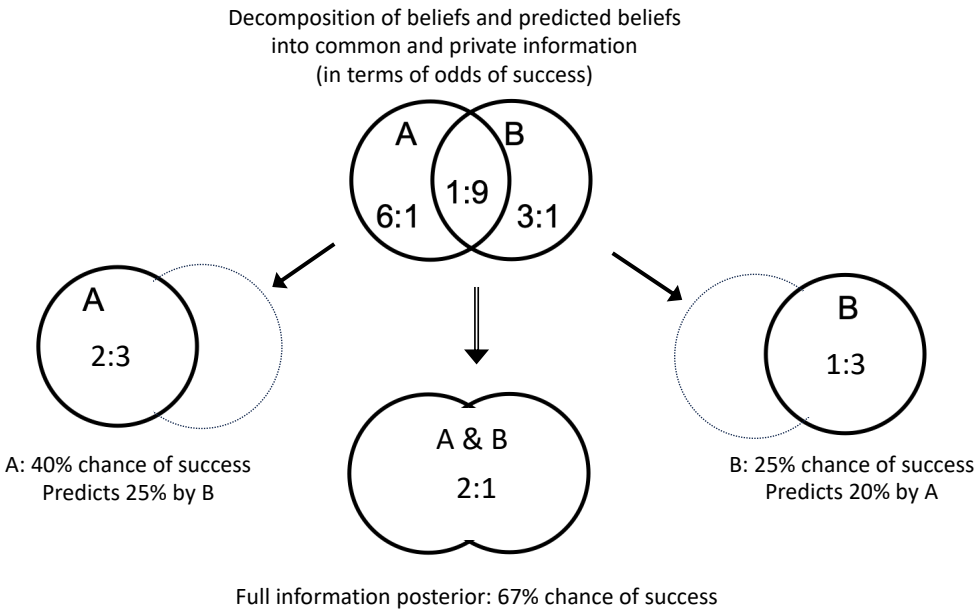

Figure I: Experts A and B have common information that suggests 1:9 odds of success, A has a private signal indicating 6:1 odds of success, and B has a private signal indicating 3:1 odds of success. This implies that A gives 2:3 odds of success (1:9 $\times$ 6:1), i.e. a 40% probability of success, and B gives 1:3 odds (1:9 $\times$ 3:1), i.e. a 25% probability of success. Given only these probability estimates from A and B, both of which suggest failure, one could not infer this decomposition. However, as the text describes, predictions from each expert about the belief of the other expert allow one to to decompose beliefs. To obtain the full information posterior under an i.i.d. signal model, the prior odds and the two likelihood ratios are multiplied giving 2:1 odds of success, or a 67% probability of success.

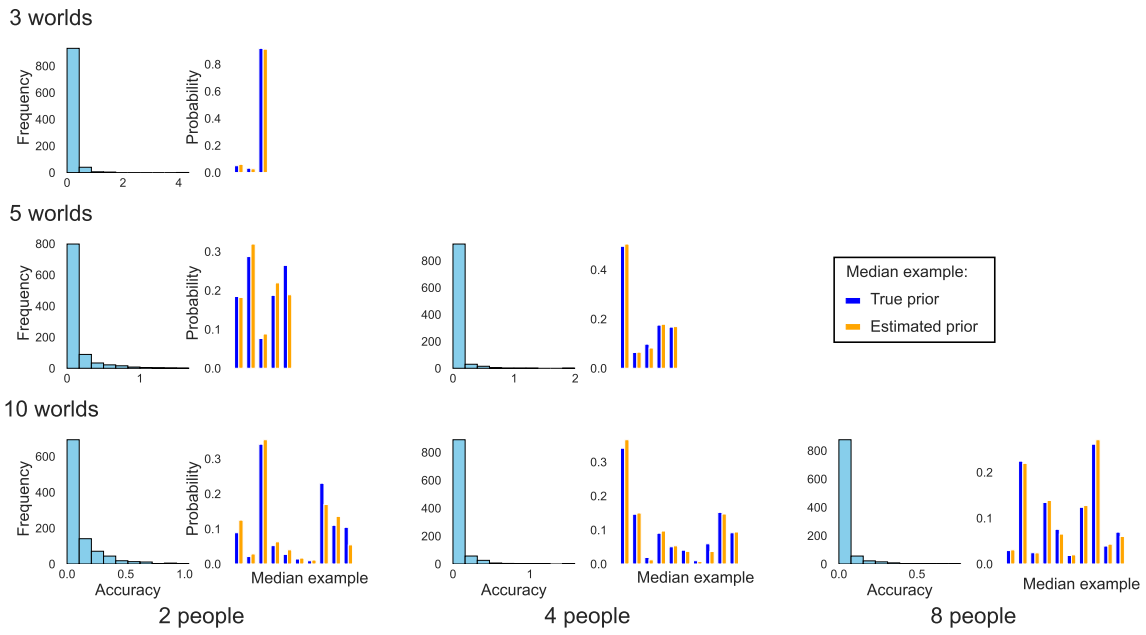

Figure II: Accuracy of ridge regression estimates of the world prior with fewer experts than possible world states. As described in the text and Supplemental Appendix, we simulate 1000 trials for each specified number of possible worlds and experts using the possible worlds model with continuous signals. We evaluate accuracy with the KL-divergence. For each number of possible worlds and experts we show a histogram of accuracies across the 1000 trials. We also show the actual and estimated prior for the trial with median accuracy to help interpret the accuracy results.

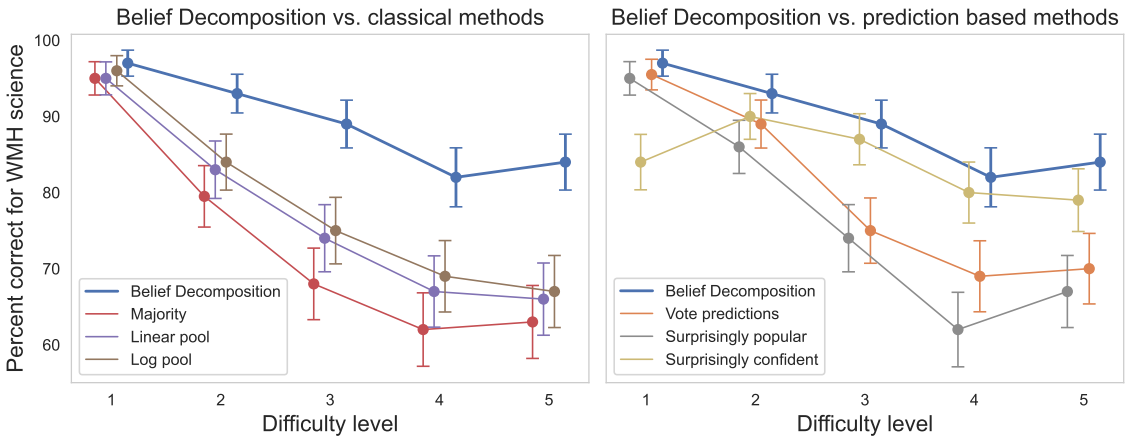

Figure III: Accuracy of aggregation methods by ex ante question difficulty for WMH science dataset. Error bars are bootstrapped standard errors.

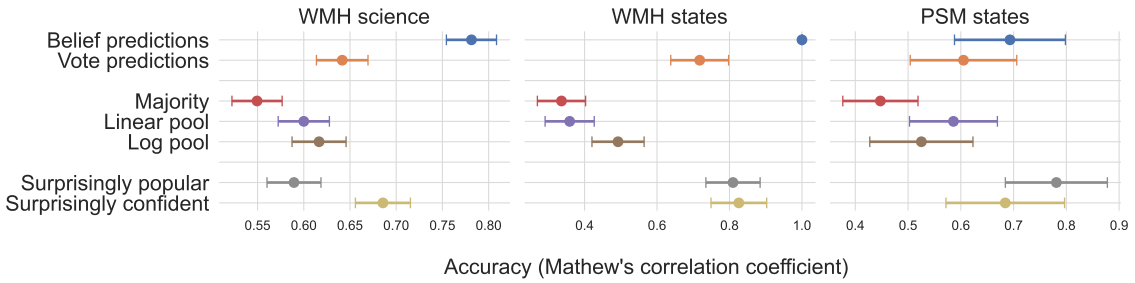

Figure IV: Categorical accuracy (Matthews correlation coefficient) of aggregation methods, including Belief Decomposition, applied to the full sample of respondents. Error bars are bootstrapped standard errors.

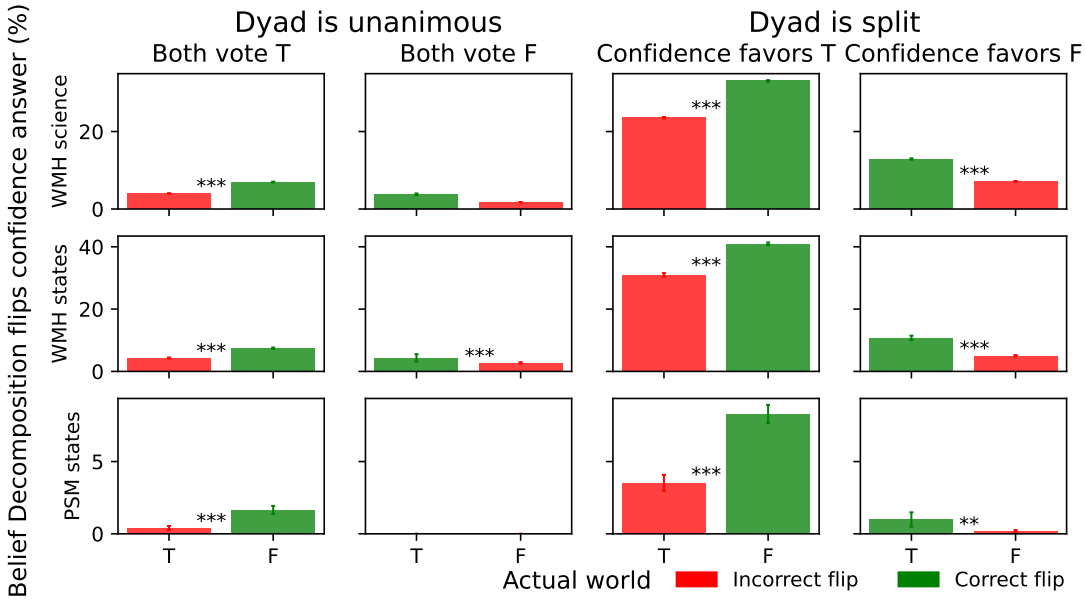

Figure V: The percentage of dyads, of different kinds, for which BD flips the confidence-based answer, conditional on the correct answer. For a given dyad kind (unanimous or split, and confidence favoring True or False), switching the confidence-based answer is correct in one world state (shown in green), and incorrect for the other (shown in red). For every dyad kind, the percentage of switches of dyads for which the switch is correct exceeds or equals the percentage of switches of dyads for which the switch is incorrect, and this difference is significant (***: $p < .0001$, **: $p < .01$) for all but two cases.

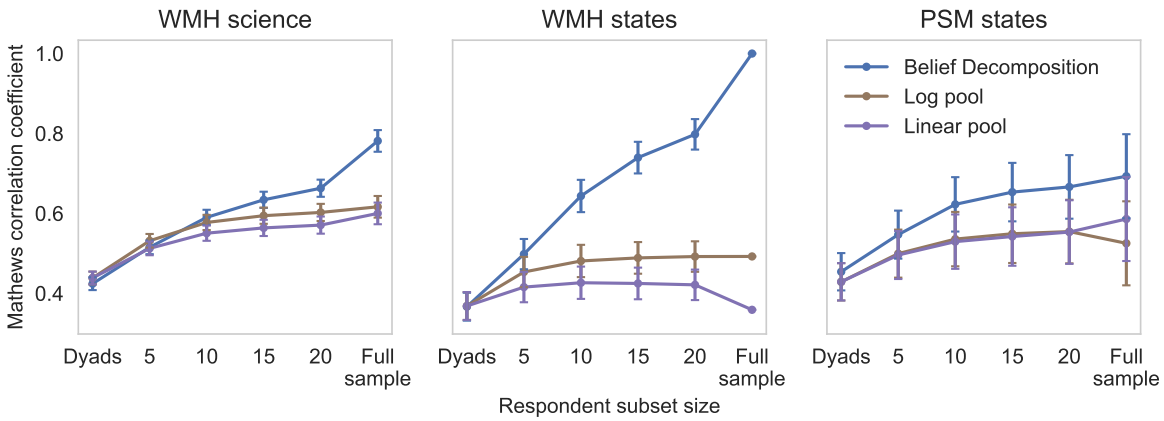

Figure VI: The accuracy of BD versus the logarithmic pool and linear pool as a function of respondent subset size. Accuracy is the average of Matthew's correlation coefficient across respondent subsets of the specified size. Error bars are bootstrapped standard errors.

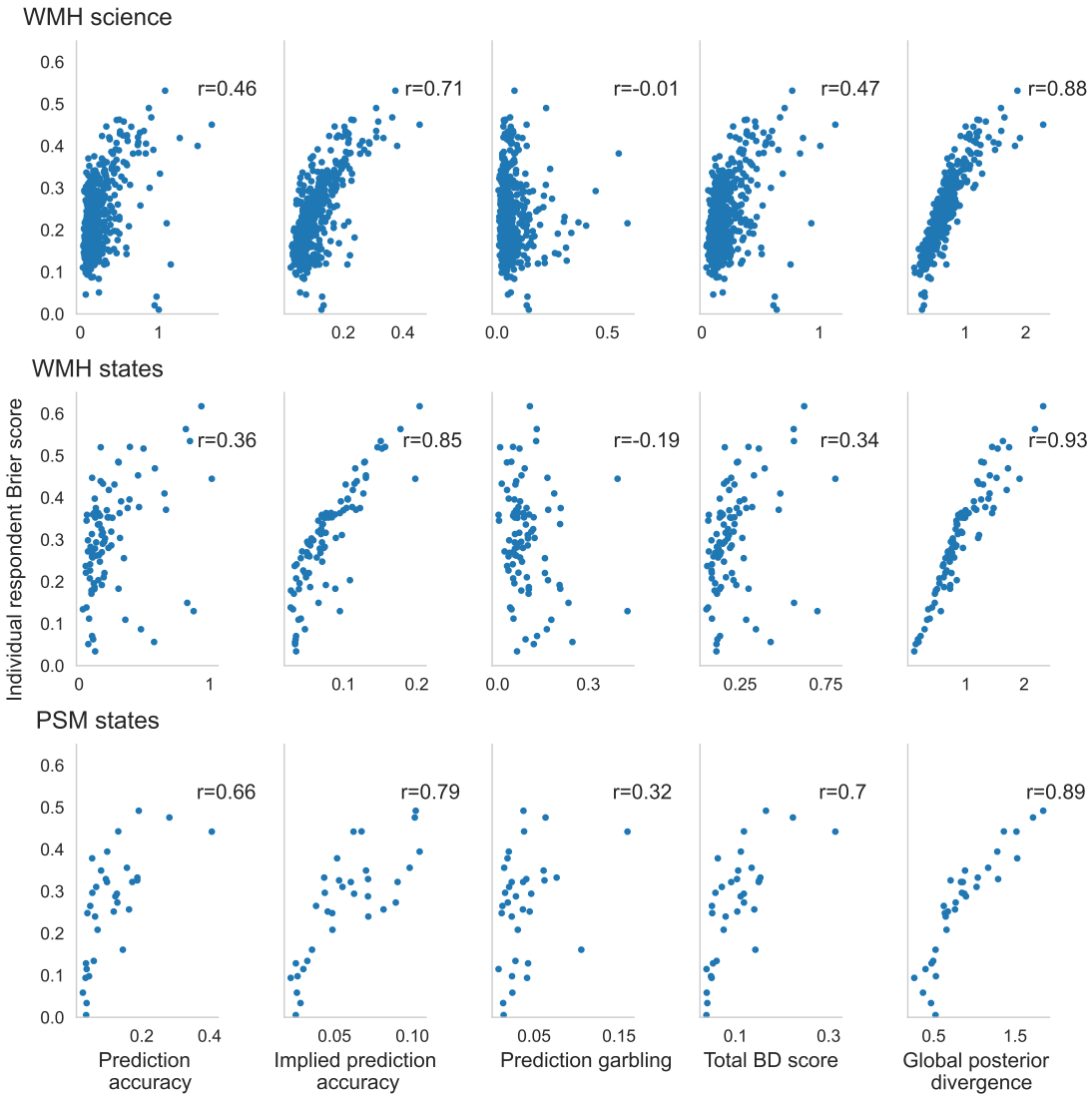

Figure VII: Respondent scores output by BD track expertise. The total score does not include side-payments and has $\alpha = .5$ and $\lambda = .5$. The full information posterior divergence is the KL divergence between respondent beliefs and the BD full information posterior. The x-axis differs across subplots.

BELIEF DECOMPOSITION
A MECHANISM FOR COLLECTIVE INFERENCE
SUPPLEMENTAL APPENDIX

John McCoy[*] and Dražen Prelec[†]

Ridge regression simulation details

We use synthetic data to validate our use of ridge regression to estimate the common prior when there are fewer experts than worlds. For a specified number of possible worlds m, beliefs and belief predictions are simulated from n respondents. We assume continuous signals in the interval $[0,1]$ and so represent world ω_i as a Beta distribution $\text{Beta}(a_i,b_i)$ where a_i and b_i are sampled uniformly from the interval $[1,10]$. The prior over worlds $p(\omega)$ is sampled from a flat Dirichlet distribution over the m-dimensional simplex. The actual world ω_{i^*} is sampled from a multinomial distribution parameterized with $p(\omega)$. The probability of a signal s conditional on a world state ω_i is thus $\text{Beta}(a_i,b_i).\text{pdf}(s)$.

Given the sampled set of worlds and prior, we compute the local expectations matrix. To compute W_{ij}, we use Monte Carlo integration to numerically estimate $W_{ij} = \int_{s^t} p(\omega_j|s^t)p(s^t|\omega_i)$. We sample 1000 signals from $\text{Beta}(a_i,b_i)$ which we denote $t_1,\ldots,t_{1000}$, and have $W_{ij} = \sum_{i=1}^{i=1000} p(\omega_j|t_i)/1000$.

For each respondent r, we sample a signal s^r from $\text{Beta}(a_{i^*},b_{i^*})$. The probability that respondent r puts on ω_i is $p(\omega_i|s^r)$, which we compute using Bayes rule. The prediction given by respondent $r$of the average probability that the sample puts on world ω_j is $\sum_j W[i,j]p(\omega_i|s^r)$.

For a fixed m and a fixed $n < m$ we repeat the above simulation for 1000 trials. For each trial, we have an $n \times m$ belief matrix B and an $n \times m$ belief prediction matrix M. Using ridge regression

[*]The Wharton School, University of Pennsylvania; jpmccoy@wharton.upenn.edu

[†]MIT, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences. E-mail: dprelec@mit.edu.

with regularization $\lambda = 0.00001$, we regress M on B to estimate a local expectation matrix $\hat{W}$ and compute the left unit eigenvector $\hat{p}$ to estimate the common prior over worlds.

We compare the estimated prior distribution $\hat{p}$ to the actual prior distribution $p(\omega)$ using KL-divergence as an accuracy metric. This ridge regression estimates the prior with only small error across different numbers of possible worlds and with only a small number of experts. Accuracy improves as the number of respondents increases and as the number of worlds decreases. For each set of worlds and number of experts, we indicate the range of accuracies over 1000 trials and compare the actual prior to the estimated prior at the median accuracy level to give an intuition for the quality of the estimate.

Further details of empirical results

Comparison of categorical accuracies on full sample (per dataset)

On the WMH science dataset, belief predictions is the most accurate method followed by the surprisingly confident answer ($M = .097\,[.050, .148]$, $z = 3.93$, $p < .001$) and is .16 MCC higher than the logarithmic pool ($M = .164\,[.105, .220]$, $z = 5.45$, $p < .001$). On the WMH states dataset, belief predictions is again the most accurate method and correctly answers every question, followed by the surprisingly confident answer ($M = .176\,[.042, .333]$,$z = 2.26$,$p = .02$). Belief predictions is approximately .5 MCC higher than the logarithmic pool ($M = .508\,[.351, .650]$, $z = 6.69$, $p < .001$) on this dataset. On the PSM states dataset, the surprisingly popular answer is most accurate, followed by belief predictions and then the surprisingly confident answer. There is no significant difference between any of these. Belief predictions is more accurate than the logarithmic pool, but not significantly ($M = .167\,[-.037, .38]$, $z = 1.57$, $p = .11$) on the PSM states dataset.

Comparison of Brier scores on full sample

Since BD formally outputs a probability, we compare it to the linear and logarithmic pool (the other two methods that output probabilities) with respect to the Brier score. The belief prediction method gives the most accurate brier score for all three datasets (WHM science: .108; WHM states: .00; PSM states: .128), but only significantly outperforms the linear and logarithmic pools for WMH science (Linear: $M = .052\,[.029, .075]$, $z = 4.31$, $p < .0001$; Logarithmic: $M = .032\,[.009, .055]$, $z = 2.72$, $p < .01$) and WMH states (Linear: $M = .207\,[.175, .240]$, $z = 12.1$, $p < .0001$; Logarithmic: $M = .185\,[.142, .231]$, $z = 8.64$, $p < .0001$).

Aggregation accuracy similarities on dyads

In principle, we wish to evaluate the accuracy of each dyad across all of the questions that the dyad answered. However, there are an intractable number of such dyads for WMH science and WMH states. Hence, for WMH science and WMH states, we construct 5000 random subsets of dyads. We construct a random subset by randomly sampling a dyad for each question. We evaluate the performance of the sample across questions, and average across samples.

BD and the confidence-based answer have similar accuracies across dyads (WMH science: $M = -.004\,[-.007, -.001]$, $z = 2.69$, $p = .007$; WMH states: $M = .005\,[-.004, .013]$, $z = 1.07$, $p = .29$; PSM states: mean difference $M = .020\,[.012, .027]$, $z = 5.10$, $p < .0001$). Although some of these differences are significant, the size of the differences are small. We also compare VP against the confidence-based answer for those dyads where the votes are split (WMH science: $M = -.003\,[-.013, .006]$, $z = 0.71$, $p = .47$; WMH states: $M = -.005\,[-.029, .02]$, $z = 0.36$, $p = .71$; PSM states: $M = .027\,[. - .003, .057]$, $z = 1.79$, $p = .07$), and again these have similar accuracies.

Pairwise method comparisons at different sample sizes

Across datasets, the accuracy of BD has a larger correlation with sample size than the accuracy of the logarithmic pool (belief predictions: $r_S = .95$, $p < .001$; logarithmic pool: $r_S = .62$, $p = .006$). For WMH science, the logarithmic pool outperforms BD for subsets of size 5 (size 5: $z = 2.65$, $p = .008$), there is no significant difference for subsets of size 10, and BD outperforms the logarithmic pool for larger subsets (size 15: $z = 2.97$, $p = .003$, size 20: $z = 3.87$, $p = .0001$). For WMH states, BD significantly outperforms the logarithmic pool for subset sizes at least five (size 5: $z = 3.11$, $p = .002$; size 10: $z = 6.76$, $p < .0001$; size 15: $z = 7.96$, $p < .0001$; size 20: $z = 8.62$, $p < .0001$). For PSM states, BD significantly outperforms the logarithmic pool across all subset sizes shown (size 5: $z = 5.33$, $p < .0001$; size 10: $z = 5.82$, $p < .0001$; size 15: $z = 4.71$, $p < .0001$; size 20: $z = 3.94$, $p < .0001$).

Comparisons between components of respondent-level scores for expertise

First, we compare the extent to which different respondent-level scores correlate with respondent expertise (Brier score). To compare, with a Steiger test for dependent correlations, the extent to which implied prediction accuracy, prediction accuarcy, and prediction garbling correlate with respondent-level Brier scores, we also require the correlation of each of these scores with the other. The cor-

relation between implied prediction accuracy and prediction accuracy is $r = .79$ for WMH science, $r = .64$ for WMH states, and $r = .54$ for PSM states. The correlation between implied prediction accuracy and prediction garbling is $r = .25$ for WMH science, $r = .20$ for WMH states, and $r = .06$ for PSM states. The correlation between prediction accuracy and prediction garbling is $r = .38$ for WMH science, $r = .61$ for WMH states, and $r = .77$ for PSM states. The correlations with each of these scores and the Brier score is given in the main text.

For WMH science, implied prediction accuracy has a significantly higher correlation with Brier score than the prediction accuracy score ($z = 11.89$, $p < .0001$) and the prediction garbling score ($z = 18.07$, $p < .0001$), and the prediction accuracy score has a significantly higher correlation with the Brier score than the prediction garbling score ($z = 9.33$, $p < .0001$). For WMH states, implied prediction accuracy has a significantly higher correlation with Brier score than the prediction accuracy score ($z = 10.56$, $p < .0001$) and the prediction garbling score ($z = 17.95$, $p < .0001$), and the prediction accuracy score has a significantly higher correlation with the Brier score than the prediction garbling score ($z = 4.54$, $p < .0001$). For PSM states, the correlation between implied prediction accuracy and the Brier score does not significantly exceed that between prediction accuracy and the Brier score ($z = 1.24$, $p = .22$), implied prediction accuracy does have a significantly higher correlation with Brier score than does prediction garbling ($z = 2.86$, $p = .008$), although this is not the case for prediction accuracy and prediction garbling ($z = 1.75$, $p = .08$).

Next, we compare the extent to which implied prediction accuracy tracks expertise compared to the Bayesian truth serum score, vote prediction accuracy, and surprisingly-confident respondent weights. We thus require the extent to which each of these correlates with implied prediction accuracy. The correlation between implied prediction accuracy and Bayesian truth serum scores is $r = .70$ for WMH science, $r = .71$ for WMH states, and $r = .82$ for PSM states. The correlation between implied prediction accuracy and vote prediction accuracy is $r = .75$ for WMH science, $r = .49$ for WMH states, and $r = .53$ for PSM states. The correlation between implied prediction accuracy and surprisingly-confident respondent weights is $r = .79$ for WMH science, $r = .58$ for WMH states, and $r = .46$ for PSM states. Implied prediction accuracy has a significantly higher correlation with Brier scores than the Bayesian truth serum scores for WMH science ($z = 13.72$, $p < .0001$) and WMH states ($z = 6.58$, $p < .0001$), but not PSM states ($z = -1.42$, $p = .17$). Implied prediction accuracy has a significantly higher correlation with Brier scores than vote prediction accuracy for WMH science ($z = 15.06$, $p < .0001$) and WMH states ($z = 12.35$, $p < .0001$), but not PSM states ($z = .90$, $p = .37$). Implied prediction accuracy has a significantly higher correlation

with Brier scores than the surprisingly-confident respondent weights for WMH science ($z = 6.58$, $p < .0001$) and WMH states ($z = 9.59$, $p < .0001$), but not PSM states ($z = .97$, $p = .34$).

Other respondent-level scores

We evaluate respondent-level scores not output by BD. An advantage of the implied prediction accuracy score and prediction garbling score is that they allow us to score probabilistic beliefs, whreas the other scores that we consider for comparison apply only to categorical answers. Since respondents provide a vote and vote prediction as well as beliefs and belief predictions we compute each respondent's Bayesian truth serum score, and, as an analogue to the prediction accuracy score, their prediction accuracy for the fraction of the sample endorsing True Prelec (2004). We also compute the weight placed on a respondent by the surprisingly-confident algorithm, which, for a given question, is simply the absolute difference between their prediction of the average belief and the actual average belief, normalized by this quantity summed across respondents Wilkening, Martinie, and Howe (2021).

The Bayesian truth serum score (WMH science: $r = .37$; all $p < .001$ unless indicated; WMH states: $r = .56$; PSM states: $r = .86$), vote prediction accuracy (WMH science: $r = .37$; WMH states: $r = .19$, $p = .07$; PSM states: $r = .70$), and the surprisingly-confident respondent weights (WMH science: $r = .56$; WMH states: $r = .35$; PSM states: $r = .69$) all correlate with Brier score performance. However, implied prediction accuracy better tracks respondent expertise than all of these for WMH science and WMH states (by a Steiger test for dependent correlations, $p < .0001$, details in appendix), and there is no significant difference for PSM states.

Robustness of incentives to strategies that manipulate belief predictions

In addition to strategies that directly perturb predictions, we examine a strategy that reverses the orginal prediction, and a strategy that substitutes a prediction of the fraction endorsing True for a prediction of the average probability. As Table 0.1 shows, strategies that manipulate predictions result in worse scores (by a paired sample t-test across respondents) for all three datasets. The only exception is that the strategy that uses vote predictions significantly improves scores for WMH science, although it significantly worsens scores for PSM states and worsens scores, although not significantly, for WMH states.

Strategy	WMH science	WMH states	PSM states
Original answers	.208	.213	.102
Perturb by 0.05	.209	.218	.105*
Perturb by 0.10	.231***	.247***	.125***
Perturb by 0.15	.266***	.284***	.158***
Perturb by 0.20	.311***	.331***	.207***
Perturb by 0.25	.363***	.383***	.270***
Reverse predictions	.819***	.610***	.285***
Use vote predictions	.199***	.214	.170***

Table 0.1: Results of strategies that manipulate the predictions of each respondent in turn. Each cell shows the average score for a strategy (across respondents and questions) on a dataset, and a comparison (using a paired sample t-test) of respondent scores under the strategy compared to honestly reporting elicited predictions. We denote $p < .05$ with *, $p < .01$ with **, and $p < .001$ with ***. We highlight in red the single case where a strategy significantly improves scores compared to the original predictions for a particular dataset.

References

Prelec, D. (2004). "A Bayesian truth serum for subjective data". *Science* 306.5695, pp. 462–466.

Wilkening, Tom, Marcellin Martinie, and Piers DL Howe (2021). "Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems". *Management Science*.