Supplementary Online Material

Here is a self-contained proof of the theorem stated in the paper. There is a countably infinite population of respondents, indexed by $r, s,.. \in I = \{1,2,\ldots\}$, who all face a common multiple-choice question, with m possible answers. The private 'signal' for respondent r is given by the m-dimensional unit vector $t^r \in E^m$, having value one on the coordinate corresponding to the true answer for respondent r and value zero on all other coordinates. A generic respondent holding opinion i is denoted by t_i (hence superscripts refer to individuals, subscripts to classes of individuals; i.e., t^r refers to the opinion of a specific respondent $r \in I$, while t_i refers to a respondent player holding opinion i).

Each respondent endorses one answer and predicts the fraction of respondents that will endorse each possible answer. The answer is represented by the n-dimensional unit vector $x^r = (x_1^r,.., x_m^r) \in E^m$, $(x_k^r \in \{0,1\}, \sum_k x_k^r = 1)$, and prediction by a relative frequency distribution $y^r = (y_1^r,.., y_m^r) \in \Delta^m$ (i.e., an element from the unit simplex in R^m, $y_k^r \geq 0$, $\sum_k y_k^r = 1$). x_k^r has value one or zero depending on whether person r has or has not endorsed answer k; y_k^r is r's estimate of the proportion of players who will endorse k. (x,y) is the (countably infinite) vector of answers and predictions.

The score of any respondent depends on his answer, his predictions, and on the empirical averages:

$$\bar{x}_k = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} x_k^r,$$

$$\log \bar{y}_k = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \log y_k^r.$$

$\bar{x}_k$ is the average frequency of answer k, and $\bar{y}_k$ the geometric average of the predicted frequencies of response k. We set $\bar{y}_k=0$ if $y_k^r=0$ for some r, and also set $\log(0/0) = 0$, and $0 \log(0) = 0$.

The score for player r, as function of (x,y), is:

$$u^r(x,y) = \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \lambda \sum_k \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}.$$

The first term, or the *information-score*, picks out $\log(\bar{x}_k/\bar{y}_k)$ that corresponds to the answer endorsed by respondent r. The second term is the *prediction score*.

<u>Information assumptions</u>

There are three information assumptions, which hold for any finite subset $S = \{r,\ldots,s\}$ of respondents,

(A1) There exists a common prior $p(t^r,\ldots,t^s)$ over opinions of member of S.

(A2) The prior is exchangeable: $p(t^r,\ldots,t^s) = p(t^{\pi(r)},\ldots,t^{\pi(s)})$, for all permutations, π, defined on the set S.

(A3) Different opinions imply different posterior distributions: $p(\bullet \mid t^r) = p(\bullet \mid t^s)$ implies $t^r = t^s$.

(A1) is a standard assumption in Bayesian game theory (*26*). (A2) is critical. By De Finetti's theorem (*25*), it implies the existence of a probability distribution $p(\theta)$ on $\Theta = \Delta^n$, which expresses the common prior as a joint distribution of conditionally independent random variables,

$$p(t^r,\ldots,t^s) = \int_{\Theta} \prod_{q \in S} p(t^q \mid \theta)\, p(\theta)\, d\theta$$

$$p(t_k^r \mid \theta) = \theta_k = \lim_{n \to \infty} \frac{1}{n} \sum_{q=1}^{n} t_k^q. \qquad (S1)$$

In other words, respondents believe that their opinions are independent, conditional on the population frequency of opinions, θ. The assumption of conditional independence may be invoked directly rather than derived from exchangeability. However, exchangeability highlights the key underlying property, that respondents with the same opinion have the same posterior beliefs about the distribution of opinions in the

population. The final assumption (A3) is a version of *stochastic relevance* (*13*); it only affects whether the truth-telling Nash equilibrium is strict.

Strategies in Bayesian Nash Equilibrium

The *answer strategy* of player r is a function $x^r(t^r) = (x_1{}^r(t^r),.., x_m{}^r(t^r)) : E^m \to \Delta^n$, indicating if player r's truthful answer is t^r, he will give answer x_k with probability $x_k{}^r(t^r)$. An answering strategy is *truthful* if $x^r(t^r) = t^r$. The *prediction strategy* of player r is likewise a function $y^r(t^r) = (y_1{}^r(t^r),.., y_m{}^r(t^r))$: $E^m \to \Delta^n$, indicating that if player r's holds opinion t^r, he will predict that a fraction $y_k{}^r$ of the population will give answer k. We don't need to consider randomized predictions, because the payoff function is strictly convex in $y_k{}^r$. The pair $(x^r(t^r),y^r(t^r))$ is then a *strategy* for player r. $(x(t),y(t))$ denotes all players' strategies, and $(x^{-r}(t^{-r}),y^{-r}(t^{-r}))$ the strategies of all players except player r. $(x(t),y(t))$ is *collectively truthful* if all answers are truthful and if predictions are consistent with Bayes' rule.

Definition $(x(t),y(t))$ is a Bayesian Nash Equilibrium (BNE) if for all players r, answers $x_k{}^r$, and predictions y^r,

$$E\{u^r(x(t),y(t)) \mid t^r\} \geq E\{u^r(x^r, y^r, x^{-r}(t^{-r}),y^{-r}(t^{-r})) \mid t^r\}, \text{ for all } x^r \in E^m,\ y^r \in \Delta^n.$$

A BNE is strict if the inequality is strict.

Theorem 1 If (A1)-(A3) hold, then collective truth-telling is a strict Bayesian Nash Equilibrium.

Proof If everyone tells the truth, then the population averages are:

$$
\begin{aligned}
\bar{x}_k &= \lim_{n\to\infty} \frac{1}{n}\sum_{s=1}^{n} x_k^s \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{s=1}^{n} t_k^s \\
&= \theta_k, \\
\log \bar{y}_k &= \lim_{n\to\infty} \frac{1}{n}\sum_{s=1}^{n} \log y_k^s \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{s=1}^{n} \log p(t_k|t^s) \\
&= \sum_{j=1}^{n} \theta_j \log p(t_k|t_j)
\end{aligned}
\tag{S2}
$$

Consider an individual with opinion i, and who believes that everyone else is truth-telling. Which answer should he endorse to maximize expected score? Because the expected prediction error does not depend on his answer, we can ignore the second part of the scoring equation, and concentrate on the information-score. If he endorses answer j, the expected information-score is:

$$E\{\, log \frac{\bar{x}_j}{\bar{y}_j} \mid t_i \,\} = \int_{\theta} p(\theta \mid t_i)\, E\{\, log \frac{\bar{x}_j}{\bar{y}_j} \mid \theta \,\}\, d\theta \qquad (S3a)$$

$$= \int_{\theta} p(\theta \mid t_i) \sum_{k=1}^{m} \theta_k \; log \frac{\theta_j}{p(t_j|t_k)}\, d\theta \qquad (S3b)$$

$$= \sum_{k=1}^{m} \int_{\theta} p(\theta, t_k \mid t_i)\, log \frac{\theta_j}{p(t_j|t_k)}\, d\theta \qquad (S3c)$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i)\, log \frac{\theta_j}{p(t_j|t_k)}\, d\theta \qquad (S3d)$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i)\, log \frac{p(t_j|\theta)\, p(t_k \mid t_j, \theta)}{p(t_j|t_k)\;\; p(t_k \mid \theta)}\, d\theta \qquad (S3e)$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i)\, log \frac{p(\theta \mid t_k, t_j)}{p(\theta \mid t_k)}\, d\theta\,. \qquad (S3f)$$

(S3a) takes the expectation of $log(\bar{x}_j / \bar{y}_j)$, with respect to the posterior distribution, $p(\theta|t_i)$, and (S3b) follows from collective truthfulness, (5). (S3c) uses conditional independence to write $\theta_k\, p(\theta|t_i)$ as $p(\theta,t_k,|t_i)$. (S3d) follows from (S3c) by Bayes' rule. In (S3e) we use conditional independence again to write θ_j as $p(t_j|\theta)$, and then to insert $p(t_k|t_j,\theta)/p(t_k|\theta) = 1$, into the fraction. (S3f) then follows by Bayes' rule.

One can now compare the expected information-score associated with truthfully endorsing answer *i* and falsely endorsing some other answer *j*:

$$E\{ln\frac{\bar{x}_i}{\bar{y}_i} \mid t_i\} - E\{ln\frac{\bar{x}_j}{\bar{y}_j} \mid t_i\} = \sum_{k=1}^{m} p(t_k \mid t_i)\int_{\Omega} p(\omega \mid t_k, t_i)\, log\frac{p(\omega \mid t_k, t_i)}{p(\omega \mid t_k)}\, d\omega$$

$$- \sum_{k=1}^{m} p(t_k \mid t_i)\int_{\Omega} p(\omega \mid t_k, t_i)\, log\frac{p(\omega \mid t_k, t_j)}{p(\omega \mid t_k)}\, d\omega$$

$$= - \sum_{k=1}^{m} p(t_k \mid t_i)\int_{\Omega} p(\omega \mid t_k, t_i)\, log\frac{p(\omega \mid t_k, t_j)}{p(\omega \mid t_k, t_i)}\, d\omega \qquad (S4)$$

$$> - \sum_{k=1}^{m} p(t_k \mid t_i)\, log\left(\int_{\Omega} p(\omega \mid t_k, t_i)\, \frac{p(\omega \mid t_k, t_j)}{p(\omega \mid t_k, t_i)}\, d\omega\right)$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i)\, log\left(\int_{\Omega} p(\omega \mid t_k, t_j)\, d\omega\right) = \sum_{k=1}^{m} p(t_k \mid t_i)\, ln\,(1) = 0$$

The inequality follows from Jensen's inequality (*29*), and is strict if (A3) holds. This proves that a truthful answer maximizes expected information-score, assuming that all other answers and predictions are truthful.

It remains to show that predictions should be truthful as well. This time we can ignore the information-score, and calculate expected prediction-score conditional on opinion *i*:

$$E\{\sum_{k=1}^{m} \bar{x}_k\, log\frac{y_k}{\bar{x}_k} \mid t_i\} = \sum_{k=1}^{m} E\{\bar{x}_k \mid t_i\}\, log\, y_k - E\{\sum_{k=1}^{m} \bar{x}_k\, log\, \bar{x}_k \mid t_i\} \qquad (S5)$$

$$= \sum_{k=1}^{m} E\{\omega_k \mid t_i\}\, log\, y_k - E\{\sum_{k=1}^{m} \omega_k\, log\, \omega_k \mid t_i\}.$$

This presumes truthful answers, so that $\bar{x}_k = \omega_k$. The second expectation does not involve y_k. Hence, the score-maximizing prediction of the proportion of *k*-answers, is the expected frequency of answer k, or, equivalently, the probability that randomly selected person holds opinion *k*: $y_k = E\{\omega_k \mid t_i\} = p(t_k \mid t_i)$.

This completes the proof that collective truth-telling is a strict Bayesian Nash Equilibrium. We now turn to,

Theorem 2 If (A1)-(A3) hold, then the following are true for all respondents:

(a) the expected information-score in any Bayesian Nash Equilibrium is non-negative,
(b) the expected information-score is (weakly) greater in the truth-telling equilibrium than in any other Bayesian Nash Equilibrium.

Proof (a) The expected information-score is obtained by setting $j=i$ in (S3):

$$E\{ \ln \frac{\bar{x}_i}{\bar{y}_i} \mid t_i \} = \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i) \ln \frac{p(\theta \mid t_k, t_i)}{p(\theta \mid t_k)} d\theta \geq 0, \qquad (S6)$$

This is the relative entropy between the distribution $p(\theta|t_k,t_i)$ and $p(\theta|t_k)$, averaged over all t_k. It measures how much knowing t_i improves other players' forecast of θ (also called "expected utility of data," regarding t_i as 'data,' Proposition 2.31 in (*25*)). The expression attains a minimum value of zero in the case where i's answer does not change others' forecasts of θ. This proves 2a.

Proof (b) We consider the joint distribution $q(\theta, \bar{x})$ induced by a particular Bayesian Nash equilibrium, $(x(t),y(t))$ (the marginal of q then coincides with p, $q(\theta)=p(\theta)$). The expected information-score for endorsing answer j when the true opinion is i equals:

$$E\{\, log \frac{\bar{x}_j}{\bar{y}_j} \mid t_i \,\} = \int_{\theta} p(\theta \mid t_i)\, E\{\, log \frac{\bar{x}_j}{\bar{y}_j} \mid \theta \,\}\, d\theta \tag{S7a}$$

$$= \int_{\theta} p(\theta \mid t_i) \sum_{k=1}^{m} \theta_k \; log \frac{\bar{x}_j}{q(x_j|t_k)}\, d\theta \tag{S7b}$$

$$= \sum_{k=1}^{m} p(t_k| t_i) \int_{\theta} p(\theta \mid t_k, t_i)\, log \frac{\bar{x}_j}{q(x_j|t_k)}\, d\theta \tag{S7c}$$

$$= \sum_{k=1}^{m} p(t_k| t_i) \int_{\bar{X}} q(\bar{x} \mid t_k, t_i)\, log \frac{\bar{x}_j}{q(x_j|t_k)}\, d\bar{x} \tag{S7d}$$

$$= \sum_{k=1}^{m} p(t_k| t_i) \int_{\bar{X}} q(\bar{x} \mid t_k, t_i)\, log \frac{q(x_j|\bar{x})\, q(t_k \mid x_j,\bar{x})}{q(x_j|t_k)\;\; q(t_k \mid \bar{x})}\, d\bar{x} \tag{S7e}$$

$$= \sum_{k=1}^{m} p(t_k| t_i) \int_{\bar{X}} q(\bar{x} \mid t_k, t_i)\, log \frac{q(\bar{x} \mid t_k, x_j)}{q(\bar{x} \mid t_k)}\, d\bar{x} \tag{S7f}$$

This is similar to (S3), except that the substitution in (S7b) no longer assumes a truth-telling equilibrium.

We now expand the relative entropy between $q(\theta, \bar{x}|t_i,t_k)$ and $q(\theta, \bar{x}|t_k)$ in two different ways, conditioning first on θ and then on $\bar{x}$:

$$\int_{\theta, \bar{X}} q(\theta, \bar{x} \mid t_k, t_i) \log \frac{q(\theta, \bar{x} \mid t_k, t_i)}{q(\theta, \bar{x} \mid t_k)} \, d\theta \, d\bar{x} =$$

$$= \int_{\theta} q(\theta \mid t_k, t_i) \log \frac{q(\theta \mid t_k, t_i)}{q(\theta \mid t_k)} \, d\theta \qquad (S8a)$$

$$+ \int_{\bar{X}, \theta} q(\bar{x} \mid \theta, t_k, t_i) \log \frac{q(\bar{x} \mid \theta, t_k, t_i)}{q(\bar{x} \mid \theta, t_k)} \, d\bar{x} \, d\theta \qquad (S8b)$$

$$= \int_{\bar{X}} q(\bar{x} \mid t_k, t_i) \log \frac{q(\bar{x} \mid t_k, t_i)}{q(\bar{x} \mid t_k)} \, d\bar{x} \qquad (S8c)$$

$$+ \int_{\theta, \bar{X}} q(\theta \mid \bar{x}, t_k, t_i) \log \frac{q(\theta \mid \bar{x}, t_k, t_i)}{q(\theta \mid \bar{x}, t_k)} \, d\theta \, d\bar{x} \qquad (S8d)$$

(S8a) only involves the marginal of $q(\theta, \bar{x})$ with respect to θ, so we can replace q with p; (S8b) equals zero, because $\bar{x}$ is fully determined by θ. Now we can write (S8a) as (S9a), and expand (S8c) into (S9b) and (S9c):

$$\int_{\theta} p(\theta \mid t_k, t_i) \log \frac{p(\theta \mid t_k, t_i)}{p(\theta \mid t_k)} \, d\theta \qquad (S9a)$$

$$= \int_{\bar{X}} q(\bar{x} \mid t_k, t_i) \log \frac{q(\bar{x} \mid t_k, x_j)}{q(\bar{x} \mid t_k)} \, d\bar{x} \qquad (S9b)$$

$$+ \int_{\bar{X}} q(\bar{x} \mid t_k, t_i) \log \frac{q(\bar{x} \mid t_k, t_i)}{q(\bar{x} \mid t_k, x_j)} \, d\bar{x} \qquad (S9c)$$

$$+ \int_{\theta, \bar{X}} q(\theta \mid \bar{x}, t_k, t_i) \log \frac{q(\theta \mid \bar{x}, t_k, t_i)}{q(\theta \mid \bar{x}, t_k)} \, d\theta \, d\bar{x} \qquad (S9d)$$

Summing up over all t_k gives an expression that relates the expected information-score in truth-telling equilibrium and expected information-score in the alternative equilibrium:

$$\sum_{k=1}^{m} p(t_k \mid t_i) \int_{\Theta} p(\theta \mid t_k, t_i) \log \frac{p(\theta \mid t_k, t_i)}{p(\theta \mid t_k)} \, d\theta \qquad (S10a)$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\bar{X}} q(\bar{x} \mid t_k, t_i) \log \frac{q(\bar{x} \mid t_k, x_j)}{q(\bar{x} \mid t_k)} \, d\bar{x} \qquad (S10b)$$

$$+ \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\bar{X}} q(\bar{x} \mid t_k, t_i) \log \frac{q(\bar{x} \mid t_k, t_i)}{q(\bar{x} \mid t_k, x_j)} \, d\bar{x} \qquad (S10c)$$

$$+ \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\Theta, \bar{X}} q(\theta \mid \bar{x}, t_k, t_i) \log \frac{q(\theta \mid \bar{x}, t_k, t_i)}{q(\theta \mid \bar{x}, t_k)} \, d\theta \, d\bar{x} \qquad (S10d)$$

(S10a) is the expected information-score in the truth-telling equilibrium; (S10b) is identical to (S7f) which is the expected information-score for endorsing j in some other equilibrium; (S10c) and (S10d) are relative entropies, hence positive. This proves that expected information-score can only decrease in a non-truth-telling equilibrium.

<u>Theorem 3</u> If (A1)-(A3) hold and $\theta=1$, then the game is zero-sum, and the total score for a respondent with opinion i in the truth-telling equilibrium equals:

$$\log p(\theta \mid t_i) + K.$$

<u>Proof</u> For $\theta=1$, the scoring equation is:

$$u^r(x,y) = \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_k \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}.$$

The average total score across all respondents than equals zero, as:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} u^r(x,y) = \lim_{n\to\infty} \frac{1}{n}\left(\sum_{r=1}^{n} \sum_{k} x_k^r \log \frac{\bar{x}_k}{\bar{y}_k}\right) + \lim_{n\to\infty} \frac{1}{n}\left(\sum_{r=1}^{n} \sum_{k} \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}\right)$$

$$= \sum_{k} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} x_k^r\right) \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_{k} \bar{x}_k \left(\lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \log \frac{y_k^r}{\bar{x}_k}\right)$$

$$= \sum_{k} \bar{x}_k \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_{k} \bar{x}_k \log \frac{\bar{y}_k}{\bar{x}_k} = 0.$$

To calculate total score in the truth-telling equilibrium, we need to supplement the expression for expected information-score (S6) with an analogous expression for expected prediction-score. Truth-telling implies that $y_k = p(t_k|t_i)$ in Equation S5:

$$E\{ \sum_{k=1}^{m} \bar{x}_k \log \frac{y_k}{\bar{x}_k} \mid t_i \} = E\{ \sum_{k=1}^{m} \theta_k \log \frac{p(t_k|t_i)}{\theta_k} \mid t_i \}, \qquad (S11)$$

$$= \int_{\theta} p(\theta \mid t_i) \sum_{k=1}^{m} \theta_k \log \frac{p(t_k|t_i)}{\theta_k} \, d\theta$$

$$= \sum_{k=1}^{m} \int_{\theta} p(\theta, t_k \mid t_i) \log \frac{p(t_k|t_i)}{\theta_k} \, d\theta$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i) \log \frac{p(t_k|t_i)}{\theta_k} \, d\theta$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i) \log \frac{p(t_k|t_i)\, p(t_k|\theta)}{p(t_k|\theta)\, p(t_k|t_i, \theta)} \, d\theta$$

$$= \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i) \log \frac{p(\theta \mid t_i)}{p(\theta \mid t_k, t_i)} \, d\theta \quad \leq \quad 0.$$

This is also a relative entropy, of the distribution $p(\theta|t_k)$ and $p(\theta|t_k,t_i)$, averaged over all t_k. It measures (negatively) the entropy reduction in i's beliefs about θ produced by learning another person's opinion k. The expected prediction-score is zero when knowledge of others' answers would not change i's beliefs about θ.

The expected <u>total</u> score in the truth-telling equilibrium now combines (S6) and (S11):

$$E\{\frac{\bar{x}_i}{\bar{y}_i} + \sum_{k=1}^{m} \bar{x}_k \log \frac{p(t_k|t_i)}{\bar{x}_k} \mid t_i\} = \sum_{k=1}^{m} p(t_k \mid t_i) \int_{\theta} p(\theta \mid t_k, t_i) \log \frac{p(\theta \mid t_i)}{p(\theta \mid t_k)} d\theta$$

$$= \int_{\theta} p(\theta \mid t_i) \sum_{k=1}^{m} \theta_k \log \frac{p(\theta \mid t_i)}{p(\theta \mid t_k)} d\theta.$$

The ex-post score for a person who endorses i, when the population average is θ, is then:

$$\sum_{k=1}^{m} \theta_k \ln \frac{p(\theta \mid t_i)}{p(\theta \mid t_k)} = \ln p(\theta \mid t_i) - \sum_{k=1}^{m} \theta_k \ln p(\theta \mid t_k),$$

The second term is a constant, determined by the zero-sum constraint. This proves Theorem 3.