A Bayesian truth serum for subjective data

Drazen Prelec

September 6, 2004

MIT, Sloan School of Management
E56-320, 38 Memorial Drive
Cambridge MA 02139
dprelec@mit.edu.

Abstract

Subjective judgments are an essential but problematic information source for science and policy — problematic, because there are no public criteria for assessing judgmental truthfulness. I present a scoring method for eliciting truthful subjective data in situations where objective truth is unknowable. The method assigns high scores, not to the most common answers, but to answers that are more common than collectively predicted, with predictions drawn from the same population. This simple adjustment in the scoring criterion removes all bias in favor of consensus: Truthful answers maximize expected score even for respondents who believe that their answer represents a minority view.

Subjective judgment, from expert and lay sources, is woven into all human knowledge. Surveys of behaviors, attitudes, and intentions are a research staple in political science, psychology, sociology and economics (*1*). Subjective expert judgment drives environmental risk analysis, business forecasts, historical inferences, artistic and legal interpretations (*2*).

The value of subjective data is limited by its quality at the source — the thought process of an individual respondent or expert. Quality would plausibly be enhanced if respondents felt as if their answers were being evaluated by an omniscient scorer, in possession of the truth (*3*). This is the situation with tests of objective knowledge, where success is defined as agreement with the scorer's answer key, or, in the case of forecasts, an observable outcome (*7*). Such evaluations are rarely appropriate in social science, because the scientist is reluctant to impose a particular definition of truth, even were one available (*8*).

Here I present a method of eliciting subjective information, designed for situations where objective truth is intrinsically or practically unknowable (*9*). The method consists of an 'information-scoring' system that induces truthful answers from a sample of rational, i.e., Bayesian, expected-value maximizing respondents. Unlike other Bayesian elicitation mechanisms (*12-14*), the method does not assume that the researcher knows the probabilistic relationship between different responses. Hence it can be applied to novel questions, by a researcher who is a complete outsider for the domain. Unlike earlier approaches to 'test theory without an answer key' (*8*), or the Delphi method (*15*), it does not privilege the consensus answer. Hence, there is no reason for respondents to bias their answer toward the likely group mean. Truthful responding remains the correct strategy even for someone who is sure that their answer represents a minority view.

The 'surprisingly common' criterion. Instead of using consensus as a truth criterion, my method assigns high scores to answers that are more common than collectively predicted, with predictions drawn from the same population that generates the answers. Such responses are surprisingly common, and the associated numerical index is called an information-score. This adjustment in the target criterion removes the bias inherent in consensus-based methods, and levels the playing field between typical and unusual opinions.

The scoring works at the level of a single question. For example, we might ask:

(i) What is your probability estimate that humanity will survive past the year 2100 (one hundred point probability scale)?

(ii) Will you vote in the next presidential elections (Definitely / Probably / Probably Not / Definitely Not)?

(iii) Have you had more than 20 sexual partners over the past year (Yes / No)?

(iv) Is Picasso your favorite twentieth-century painter (Yes / No)?

Each respondent provides a personal answer and also a prediction of the empirical distribution of answers, i.e., the fraction of people endorsing each answer. Predictions are scored for accuracy — for how well they match the empirical frequencies. The personal answers, which are the main object of interest, are scored for being surprisingly common. An answer endorsed by 10% of the population against a predicted frequency of 5% would be surprisingly common and would receive a high information score; it would be a surprisingly *un*common, hence low scoring, answer if predictions averaged 25%.

The surprisingly common criterion exploits an overlooked implication of Bayesian reasoning about population frequencies, namely, that in most situations one should expect that others will underestimate the true frequency of one's own opinion or personal characteristic. This implication is a corollary to the more usual Bayesian argument that the highest predictions of the frequency of a given opinion or characteristic in the population should come from individuals who hold that opinion or characteristic, because holding the opinion constitutes a valid and favorable signal about its general popularity (*16, 17*). People who, for example, rate Picasso as their favorite should — and usually do *(18)* — give higher estimates of the percentage of the population who shares that opinion, because their own feelings are an informative 'sample of one' (*21*). It follows, then, that Picasso lovers — who have reason to believe that their best estimate of Picasso popularity is high compared to others' estimates — should conclude that the true popularity of Picasso is underestimated by the population. Hence, one's true opinion is also the opinion that has the best chance of being surprisingly common.

The validity of this conclusion does not depend on whether the personally truthful answer is believed to be rare or widely shared. For example, a male who has had more than 20 sexual partners [answering question (iii)] may feel that few people fall in this promiscuous category. Nevertheless, according to Bayesian reasoning, he should expect that his personal estimate of the percentage (e.g., 5%) will be somewhat higher than the average of estimates collected from the population as a whole (e.g., 2%). The fact that he has had more than 20 sexual partners is evidence that the general population — which includes persons with fewer partners — will underestimate the prevalence of this profile.

Truth-telling is individually and collectively optimal. Truth-telling is individually rational in the sense that a truthful answer maximizes expected information-

score, assuming that everyone is responding truthfully [hence it is a Bayesian Nash equilibrium (*23*)]. It is also collectively rational in the sense that no other equilibrium provides a higher expected information-score, for any respondent. In actual applications of the method, one would not teach respondents the mathematics of scoring or explain the notion of equilibrium. Rather, one would like to be able to tell them that truthful answers will maximize their expected scores, and that in arriving at their personal true answer they are free to ignore what other respondents might say. The equilibrium analysis confirms that under certain conditions one can make such a claim honestly.

The equilibrium results rest on two assumptions. First, the sample of respondents is sufficiently large so that a single answer cannot appreciably affect empirical frequencies (*24*). The results do hold for large finite populations but are simpler to state for a countably infinite population, as is done here. Respondents are indexed by $r \in \{1,2,\ldots\}$, and their truthful answer to a m-multiple choice question by $t^r = (t_1^r,.., t_m^r)$ ($t_k^r \in \{0,1\}$, $\sum_k t_k^r = 1$). t_k^r is thus an indicator variable having value one or zero depending on whether answer k is or is not the truthful answer of respondent r. The truthful answer is also called a personal opinion or characteristic.

Second, respondents treat personal opinions as an 'impersonally informative' signal about the population distribution, which is an unknown parameter, $\theta = (\theta_1,..,\theta_m) \in \Theta$ *(25).* Formally, I assume common knowledge (*26*) by respondents that all posterior beliefs, $p(\theta \mid t^r)$, are consistent with Bayesian updating from a single distribution over θ, also called a common prior, $p(\theta)$, and that: $p(\theta \mid t^r) = p(\theta \mid t^s)$ if and only if $t^r = t^s$. Opinions thus provide evidence about θ but the inference is impersonal: respondents believe that others sharing their opinion will draw the same inference about population frequencies (*27*). One can therefore denote a generic respondent with opinion j by t_j, and suppress the respondent superscript from joint and conditional probabilities: $Prob\{ t_j^r = 1 \mid t_i^s = 1\}$ becomes $p(t_j|t_i)$, etc..

For a binary question one may interpret the model as follows. Each respondent privately and independently conducts one toss of a biased coin, with unknown probability θ_H of heads. The result of the toss represents his opinion. Using this datum, he forms a posterior distribution, $p(\theta_H \mid t^r)$, whose expectation is the predicted frequency of Heads. For example, if the prior is uniform, then the posterior distribution following the toss will be triangular on [0,1], skewed toward Heads or Tails depending on the result of the toss, with expected value of 1/3 or 2/3. However, if the prior is not uniform but strongly biased toward the opposite result, i.e., Tails, then the expected frequency of Heads following a Heads toss might still be quite low. This would correspond to a prima facie unusual characteristic, such as having more than 20 sexual partners within the previous year.

An important simplification in the method is that I never elicit prior or posterior distributions, only answers and predicted frequencies. Denoting answers and predictions by $x^r = (x_1^r,.., x_m^r)$ $(x_k^r \in \{0,1\}, \sum_k x_k^r = 1)$, and $y^r = (y_1^r,.., y_m^r)$ $(y_k^s \geq 0, \sum_k y_k^s = 1)$, respectively, I calculate the population endorsement frequencies, $\bar{x}_k$, and the (geometric) average, $\bar{y}_k$, of predicted frequencies,

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

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

Instead of applying a preset answer key, we evaluate answers according to their information-score, which is the log-ratio of actual-to-predicted endorsement frequencies:

$$\text{Information-score for answer } k = \log \frac{\bar{x}_k}{\bar{y}_k} \tag{1}$$

At least one answer will have a non-negative information-score. Variance in predictions tends to lower all $\bar{y}_k$ and hence raises information-scores.

The total score for a respondent combines the information-score with a separate score for the accuracy of predictions (*28*):

$$\begin{aligned} \text{Score for respondent } r &= \text{Information-score + Prediction score} \\ &= \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \alpha \sum_k \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}, \qquad 0 < \alpha \leq 1. \end{aligned} \tag{2}$$

Equation 2 is the complete payoff equation for the game. It is symmetric, and zero sum if $\alpha=1$. The first part of the equation selects a single information-score value, as $x_k^r=0$ for all answers except the one endorsed by *r*. The second part is a penalty proportional to the relative entropy (or Kullback-Leibler divergence) between the empirical distribution and *r*'s prediction of that distribution (*29, 30*). The best prediction score is zero, attained when prediction exactly matches reality, $y_k^r = \bar{x}_k$. Expected prediction score is maximized by reporting expected frequencies, $y_k^r = E\{\bar{x}_k | t^r\}$ (*2*). The constant α fine-tunes the weight given to prediction error.

To see how this works in the simple coin-toss setting, imagine that there are only two equally likely possibilities — either the coin is fair, or it is unfair in which case it always comes up Heads. A respondent who privately observes a single toss of Tails knows that the coin is fair, and predicts a 50-50 split of observations. A respondent

observing Heads lowers the probability of fairness from the prior 1/2 to a posterior of 1/3, in accord with Bayes' rule, which in turn yields a predicted (i.e., expected) frequency of 1/6 for Tails (multiplying 1/3 by 1/2). From the perspective of someone observing Tails, the expectation of others' predictions of the frequency of Tails will be a mix of predictions of 1/2 (from those tossing Tails) and 1/6 (from those tossing Heads), yielding a geometric mean clearly lower than her predicted frequency of 1/2. Hence, she expects that Tails will prove to be more common than predicted and receive a positive information-score. By contrast, Heads is expected to be a surprisingly uncommon toss, because the predicted frequency of 1/2 is lower than the expectation of others' predictions, which is a mix of 1/2 and 5/6 predictions. A similar argument would show that those who draw Heads should expect that Heads will prove to be the answer with the high information-score.

The example illustrates a general property of information-scores, namely, that a truthful answer constitutes the best guess about the most surprisingly common answer, if "best" is defined precisely by expected information-score, and if other respondents are answering truthfully and giving truthful predicted frequencies. This property does not depend on the number of possible answers or the prior (*31*). It leads directly to the equilibrium result (proof in the Supporting Online Material):

> Theorem Assume that: (i) every respondent r with opinion t^r forms a posterior over the population distribution of opinions, $p(\cdot|t^r)$, by applying Bayes' rule to a common prior $p(\cdot)$; (ii) $p(\cdot|t^r) = p(\cdot|t^s)$ if and only if $t^r = t^s$; and (iii) scores are computed according to equation 2. Then,
>
> (T1) Truth-telling is a Nash equilibrium for any $\alpha>0$: Truth-telling maximizes expected total score of every respondent who believes that others are responding truthfully;
>
> (T2) Expected equilibrium information-scores are non-negative, and attain a maximum for all respondents in the truth-telling equilibrium;
>
> (T3) For $\alpha=1$, the game is zero-sum, and the total scores in the truth-telling equilibrium equal: $\log p(\cdot|t^r) + K$, with K set by the zero-sum constraint.

Truth-telling is defined as truthful answers, $x^r = t^r$, and truthful predictions, $y^r = E_p\{\cdot|t^r\}$. T2 states that although there are other equilibria, constructed by mapping multiple true opinions into a single response category or by randomization, these less revealing equilibria result in lower information-scores for all respondents. If needed, one can

enhance the strategic advantage of truth-telling by giving relatively more weight to information-score in equation 1 (*32*). For sufficiently small α, the expected *total* scores in the truth-telling equilibrium will Pareto-dominate expected scores in any other equilibrium. T3 shows that by setting $\alpha=1$ we also have the option of presenting the survey as a purely competitive, zero-sum contest. Total scores then rank respondents according to how well they anticipate the true distribution of answers. Note that the scoring system asks only for the expected distribution of true answers, $E_p\{\omega \mid t^r\}$ and not for the posterior distribution $p(\omega \mid t^r)$, which is an m-dimensional probability density function. Remarkably, one can infer which respondents assign more probability to the actual value of ω by means of a procedure that does not elicit these probabilities directly.

Respondents can freely ignore how others will answer. In previous economic research on incentive mechanisms it has been standard to assume that the scorer (or the 'center') knows the prior and posteriors and incorporates this knowledge into the scoring function (*12-14, 33*). In principle, any change in the prior, whether caused by a change in question wording, in the composition of the sample, or by new public information, would require a recalculation of the scoring functions. By contrast, my method employs a universal 'one-size-fits-all' scoring equation, which makes no mention of prior or posterior probabilities. This has three benefits for practical application. First, questions do not need to be limited to some pre-tested set for which empirically estimated base rates and conditional probabilities are available; instead, one can use the full resources of natural language to tailor a new set of questions for each application. Second, it is possible to apply the same survey to different populations, or in a dynamic setting (which is relevant to political polling). Third, one can honestly instruct respondents to refrain from speculating about the answers of others while formulating their own answer. Truthful answers are optimal for any prior, and there are no posted probabilities for them to consider, and perhaps reject.

These are decisive advantages when it comes to scoring complex, unique questions. In particular, one can apply the method to elicit honest probabilistic judgments about the truth-value of any clearly stated proposition, even if actual truth is beyond reach and no prior is available. For example, a recent book, Our Final Century, by a noted British astronomer, gives the chances of human survival beyond the year 2100 at no better than 50:50 (*34*). It is a provocative assessment, which will not be put to the test anytime soon. With the present method, one could take the question: "Is this our final century?" and submit it to a sample of experts, who would each provide a subjective probability and also estimate probability distributions over others' probabilities. T1 implies that honest reporting of subjective probabilities would maximize expected

information-score. Experts would face comparable truth-telling incentives as if they were betting on the actual outcome, e.g., as in a futures market (*10*), and that outcome could be determined in time for scoring.

I illustrate this with a discrete computation, which assumes that probabilities are elicited at 1% precision via a hundred-point multiple-choice question (in practice, one would have fewer categories, and smooth out the empirical frequencies). The population vector $\theta = (\theta_{00},..,\theta_{99})$ indexes the unknown distribution of such probabilities among experts. Given any prior, $p(\theta)$, it is a laborious but straightforward exercise to calculate expected information-score as function of true personal probability and endorsed probability. Fig. 1, lines A90 and B90, present the result of such calculations, with two different priors $p_A(\theta)$, and $p_B(\theta)$, for experts who happen to agree that the probability of disaster striking before 2100 is 90%. The experts thus share the same assessment but have different theories about how their assessment is related to the assessment of others. Although lines A90 and B90 differ, the expected information-score is in both cases maximized by a truthful endorsement of 90%. This confirms T1. In both cases, each expert believes that his subjective probability is pessimistic relative to the population: The expectation of others' probabilities, conditioned on a personal estimate of 90%, is only 65% with $p_A(\theta)$, and 54% with $p_B(\theta)$.

If the subjective probability shifts to 50%, the lines move to A50, B50, and the optimum, in both cases, relocates to 50%. Hence, the optimum automatically tracks changes in subjective belief — in this case the subjective probability of an unknown future event — but is invariant with respect to assumptions about how that belief is related to beliefs of other individuals. Changing these assumptions will simply lead back to the same recommendation — to truthfully report subjective probability.

Respondents are thus free to concentrate on their personal answer and need not worry about formulating an adequate prior. Any model of the prior is likely to be complex and involve strong assumptions. For example, in the calculations in Fig. 1, I assumed that experts' estimates are based on a private signal, distributed between zero and one, representing a personal assessment of the credibility of evidence supporting the bad outcome. The 'credibility signal' is a valid, but stochastic indicator of the true state of affairs: On the bad scenario, credibility signals are independent draws from a uniform distribution, so that some experts 'get the message' and some do not; on the good scenario, they are independent draws from a triangular distribution, peaking at zero (no credibility) and declining linearly to one (full credibility). A prior probability of catastrophe then induces a monotonic mapping from credibility signals to posterior probabilities of catastrophe, as well as a prior over experts' probability estimates, $p(\theta)$.

Lines A and B differ in that the prior probability of catastrophe is presumed to be 50% for line A, and 20% for line B. Expected scores are higher for B, because the 90% estimate is more surprising in that case.

One could question any of the assumptions of this model (*35*). However, changing the assumptions would not move the optimum, as long as the impersonally informative requirement is preserved. (The impersonally informative requirement means that two experts will estimate the same probability of catastrophe if and only if they share the same posterior distribution over other experts' probabilities). Thus, even though information-scoring conditions success on the answers of other people, the respondent does not need to develop a theory of other people's answers — the most popular answer has no advantage at being the 'winning one,' and the entire structure of mutual beliefs, as embodied in the prior, is irrelevant.

Proper scoring of probabilities. It is instructive to compare information-scores with scores that would be computed if the Scorer had a crystal ball, and could score estimates for accuracy. The standard instrument for eliciting honest probabilities about publicly verifiable events is the logarithmic proper scoring rule (*2, 7, 11*). With the rule, an expert who announces a probability distribution $z = (z_1,.., z_n)$ over n mutually exclusive events would receive a score of,

$$K + \log z_i \qquad (3)$$

if event i is realized. For instance, an expert whose true subjective probability estimate that humanity will perish by 2100 is *90%* but who announced a possibly different probability z, would calculate an expected score of $0.9 \log z + 0.1 \log(1-z)$, assuming, again, that there was some way to establish the true outcome. This expectation is maximized at the true value, $z=0.90$, as shown by line PS90 in Fig. 1 (elevation is arbitrary). It is hard to distinguish proper scoring, which requires knowledge of the true outcome, from information-scoring, which does not require such knowledge (*36*).

Informational boundary conditions. There are two generic ways in which the assumption of an impersonally informative prior might fail. First, a true answer might not be informative about population frequencies in the presence of public information about these frequencies (inducing a sharp prior). For instance, a person's gender would have minimal impact on their judgment of the proportion of men and women in the population. This would be a case of $t^r \neq t^s$ but $p(\cdot \mid t^r) \approx p(\cdot \mid t^s)$, and the difference between expected information-scores for honest and deceptive answers would be

virtually zero (though still positive). As shown below, the remedy is to combine the gender-question with an opinion question that interacts with gender.

Second, respondents with different tastes or characteristics might choose the same answer for different reasons, and hence form different posteriors. For example, someone with nonstandard political views might treat her liking for a candidate as evidence that most people will prefer someone else. This would be a case of : $p(\theta \mid t^r) \neq p(\theta \mid t^s)$ although $t^r = t^s$. Here, too, the remedy is to expand the questionnaire, allowing the person to reveal both the opinion and characteristic.

A last example, an art evaluation, illustrates both remedies. The example assumes existence of experts and laymen, and a binary state-of-nature — that a particular artist either does or does not represent an original talent. By hypothesis, art experts recognize this distinction quite well, but laymen discriminate poorly and, indeed, have a higher chance of enjoying a derivative artist than an original one. The fraction of experts is common knowledge, as are the other probabilities (given in Table 1).

In the Short Version of the survey, respondents only state their opinion; in the Long Version, they also report their expertise. Table 1 displays expected information scores for all possible answers, as function of opinion and expertise. With the Short Version, truth-telling is optimal for experts but not for laymen, who do have a slight incentive to deceive if they happen to like the exhibition. With the Long Version, however, the diagonal, truth-telling entries have highest expected score. In particular, respondents will do better if they reveal their true expertise even though the distribution of expertise in the surveyed population is common knowledge.

Experts receive higher scores. Expected information-scores in this, and other examples, reflect the amount of information associated with a particular opinion or characteristic. In Table 1 experts have a clear advantage even though they comprise a minority of the sample, because their opinion is more informative about population frequencies. In general, the expected information-score for opinion i equals the expected relative entropy between distribution $p(\theta \mid t_k, t_i)$ and $p(\theta \mid t_k)$, averaged over all t_k. In words, the expected score for i is the information-theoretic measure of how much endorsing opinion i shifts others' posterior beliefs about the population distribution. An expert endorsement will cause greater shift in beliefs, because it is more informative about the underlying variables that drive opinions for both segments (*37*). This measure of impact is quite insensitive to the size of the expert segment, or to the direction of association between expert and non-expert opinion.

By establishing truth-telling incentives, I do not suggest that people are deceitful or unwilling to provide information without explicit financial payoffs. The concern,

rather, is that the absence of external criteria can promote self-deception and false-confidence even among the well-intentioned. A futurist, or an art critic, can comfortably spend a lifetime making judgments without the reality checks that confront a doctor, scientist, or business investor. In the absence of reality checks, it is tempting to grant special status to the prevailing consensus. The benefit of explicit scoring is precisely to counteract informal pressures to agree (or perhaps to 'stand out' and disagree). Indeed, the mere existence of a truth-inducing scoring system provides methodological reassurance for social science, showing that subjective data can — if needed — be elicited via a process that is neither faith-based ('all answers are equally good') nor biased against the exceptional view.(*39*)

Endnotes

1. C. F. Turner, E. Martin, Eds., *Surveying Subjective Phenomena*, vol. I - II (Russell Sage Foundation, New York, 1984).
2. R. M. Cooke, *Experts in Uncertainty* (Oxford University Press, New York, 1991).
3. A formalized scoring rule has diverse uses: for training, as in psychophysical experiments (*4*), for communicating desired performance (*5*), for enhancing motivation and effort (*6*), for encouraging advance preparation, as in educational testing, for attracting a larger and more representative pool of respondents, for diagnosing suboptimal judgments (*7*), for identifying superior respondents.
4. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophyics* (Peninsula Publishing, Los Altos, 1989).
5. W. Edwards, *Psychol Rev* **68**, 275 (1961).
6. C. F. Camerer, R. Hogarth, *J Risk Uncert* **18**, 7 (1999).
7. R. Winkler, *J Am Stat Assoc* **64**, 1073 (1969).
8. W. H. Batchelder, A. K. Romney, *Psychometrika* **53**, 71 (1988).
9. In particular, this precludes the application of a futures markets (*10*) or a proper scoring rule (*11*).
10. J. Berg, R. Forsythe, T. A. Rietz, in *Understanding Strategic Interaction: Essays in the Honor of Reinhard Selten* W. Albers, W. Guth, B. Hammerstein, B. Moldovanu, E. van Damme, Eds. (Springer, New York, 1996).
11. L. J. Savage, *J Am Stat Assoc* **66**, 783 (1971).
12. C. d'Aspremont, L.-A. Gerard-Varet, *J Public Econ* **11**, 25 (1979).
13. S. J. Johnson, J. Pratt, R. J. Zeckhauser, *Econometrica* **58**, 873 (1990).
14. P. McAfee, P. Reny, *Econometrica* **60**, 395 (1992).
15. H. A. Linstone, M. Turoff, *The Delphi Method: Techniques and Applications* (Addison-Wesley, Reading, MA, 1975).
16. R. M. Dawes, in *Insights in Decision Making* R. Hogarth, Ed. (1990) pp. 179-199.
17. S. J. Hoch, *J Pers Soc Psychol* **53**, 221 (1987).
18. It is one of the most robust findings in experimental psychology that subjects' self-reported characteristics — behavioral intentions, preferences, beliefs — are positively correlated with their estimates of the relative frequency of these characteristics (*19*). The psychological literature initially regarded this as an egocentric error of judgment (a 'false consensus') (*20*) and did not consider the Bayesian explanation, as was pointed out by Dawes (*16, 21*). There is still some dispute whether the relationship is entirely consistent with Bayesian updating (*22*).
19. G. Marks, N. Miller, *Psychol Bull* **102**, 72 (1987).
20. L. Ross, D. Greene, P. House, *J Exp Soc Psychol* **13**, 279 (1977).
21. R. M. Dawes, *J Exp Soc Psychol* **25**, 1 (1989).
22. J. Krueger, R. W. Clement, *J Pers Soc Psychol* **67**, 596 (1994).
23. D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, MA, 2000).
24. With finite players, the truth-telling result holds provided the number of players exceeds some finite n, which in turn depends on $p(\theta)$.

25. J. M. Bernardo, A. F. M. Smith, *Bayesian Theory*, Wiley Series in Probability and Statistics (Wiley, New York, 2000).
26. R. J. Aumann, *Econometrica* **55**, 1 (1987).
27. More precisely, I assume in the Supporting Online Material that for any finite subset of respondents there is a common and exchangeable prior over their opinions (hence the prior is invariant under permutation of respondents). By de Finetti's representation theorem *(25)*, this implies the existence of a probability distribution, $p(\theta)$, such that opinions are independent conditional on θ. Conditional independence ensures $t^r = t^s \Rightarrow p(\theta \mid t^r) = p(\theta \mid t^s)$. The reverse implication, i.e., that different opinions imply different posteriors is also called stochastic relevance (*13*)
28. The finite n-player scoring formula ($n \geq 3$), for respondent r, is

$$\sum_{s \neq r} \sum_{k} x_k^r \log \frac{\bar{x}_k^{-rs}}{\bar{y}_k^{-rs}} + \alpha \sum_{s \neq r} \sum_{k} \bar{x}_k^{-rs} \log \frac{y_k^r}{\bar{x}_k^{-rs}},$$

where: $\bar{x}_k^{-rs} = \left(\sum_{q \neq r,s} x_k^q + 1 \right)/(n + m - 2)$, $\log \bar{y}_k^{-rs} = \sum_{q \neq r,s} \log y_k^q /(n - 2)$.

The score for r is built up from pairwise comparisons of r against all other respondents s, excluding from the pairwise calculations the answers and predictions of respondents r and s. To prevent infinite scores associated with zero frequencies, I replace the empirical frequencies with Laplace estimates derived from these frequencies. This is equivalent to 'seeding' the empirical sample with one extra answer for each possible choice. Any distortion in incentives can be made arbitrarily small by increasing the number of respondents, n. The scoring is zero-sum when $\alpha = 1$.
29. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
30. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1954).
31. The key step in the proof involves calculation of expected information-score for someone with personal opinion i but endorsing a possibly different answer j,

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

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

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

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

Once we reach (d), we can use the fact that the integral,

$$\int_{\theta} p(\theta \mid t_k, t_i)\ log\ p(\theta \mid t_k, t_j)\ d\theta,$$

is maximized when: $p(\theta|t_k,t_j) = p(\theta|t_k,t_i)$, to conclude that a truthful answer, *i*, will have higher expected information-score than any other answer *j*. To derive (d), we first compute expected information-score (a) with respect to the posterior distribution, $p(\theta|t_i)$, and use the assumption that others are responding truthfully to derive (b). For an infinite sample, truthful answers imply: $\bar{x}_j=\theta_j$, and truthful predictions: $log\ \bar{y}_j = \sum_k \theta_k\ log\ p(t_j|t_k)$, because the fraction θ_k of respondents who draw *k* will predict $p(t_j|t_k)$ for answer *j*. To derive (c) from (b), we apply conditional independence to write $\theta_k\ p(\theta|t_i)$ as $p(t_k|t_i)p(\theta|t_k,t_i)$, θ_j as $p(t_j|\theta)$, and *1* as $p(t_k|t_j,\theta)/p(t_k|\theta)$, which is inserted into the fraction. (d) follows from (c) by Bayes' rule.

32. The scoring system is not easy to circumvent by collective collusion, because if everyone agrees to give the same response then that response will no longer be surprisingly common, and will receive a zero information-score. The prediction scores will also be zero in that case.
33. J. Cremer, R. P. McLean, *Econometrica* **56**, 1247 (1988).
34. M. Rees, *Our Final Century* (Heinemann, London, 2003).
35. Certainly, the assumption of independent credibility signals is unrealistic in that it implies that expert opinion can in aggregate predict the true outcome perfectly; a more realistic model would have to interpose some uncertainty between the outcome and the totality of expert opinion.
36. Information scoring is non-metric, and the hundred probability levels are treated as simply one hundred distinct response categories. The smooth lines in Fig. 1

reflect smooth underlying priors, $p_A(\theta)$ and $p_B(\theta)$. Unlike proper-scoring, information-scoring could be applied to verbal expressions of probability ('likely', 'impossible,' etc.).
37. Precisely, if the opinions of one type of respondent are a statistical 'garbling' of the opinions of a second type, then the first type will receive a lower score in the truth-telling equilibrium. Garbling means that the more informed individual could replicate the statistical properties of the signal received by the less informed individual, simply by applying a randomization device to his own signal (*38*).
38. D. Blackwell, *Ann Math Stat* **24**, 265 (1953).
39. I thank Mijovic-Prelec, Frederick, Fudenberg, Hauser, Kearns, Kugelberg, Luce, McAdams, Seung, Weaver and Wernerfelt for comments and criticism. I acknowledge early support for this research direction by Harvard Society of Fellows, MIT EBusiness Center, and the MIT Center for Innovation in Product Development.

Figure and Table Captions

For Figure 1

Figure 1 illustrates how expected information-score is maximized by a truthful report of subjective belief in a proposition (i.e., that 'this is our final century'), irrespective of priors (A or B) or subjective probability values (50% or 90%). Line A90 gives expected score for different reported probabilities when true personal estimate of catastrophe is 90% and prior probability is 50%. It is optimal to report 90% even though that is expected to be an unusually pessimistic estimate. Changing the prior to 20% (line B90) increases expected scores but does not displace the optimum. Changing subjective probability to 50% shifts the optimum to 50% (A50 assumes a 50% prior, B50 a 20% prior). Standard proper scoring (expectation of eq. 3, displayed as line PS90) also maximally rewards a truthful report (90%). However, proper scoring requires knowledge of the true outcome, which may remain moot until 2100.

For Table 1

An incomplete question can create incentives for misrepresentation. The first pair of columns give the conditional probabilities of liking the exhibition as function of originality (so that, for example, Experts have a 70% chance of liking an

original artist). It is common knowledge that 25% of the sample are Experts, and that the prior probability of an original exhibition is 25%. The remaining columns display expected information-scores. Truth-telling is optimal in the Long Version, but not in the Short Version of the survey.

		Probability of opinion conditional on quality of exhibition		Expected score for reported opinion and expertise level LONG VERSION				SHORT VERSION	
		If original	If derivative	Expert claims Like	Expert claims Dislike	Layman claims Like	Layman claims Dislike	Like	Dislike
Expert opinion	Like	70%	10%	**+575**	-776	-462	+67	**+191**	-57
	Dislike	30%	90%	-934	**+95**	+84	-24	-86	**+18**
Layman opinion	Like	10%	20%	-826	+32	**+45**	-18	-66	**+12**
	Dislike	90%	80%	-499	-156	-73	**+2**	-6	**-4**

Table 1