LETTER

doi:10.1038/nature21054

A solution to the single-question crowd wisdom problem

Dražen Prelec[1,2,3], H. Sebastian Seung[4] & John McCoy[3]

Once considered provocative[1], the notion that the wisdom of the crowd is superior to any individual has become itself a piece of crowd wisdom, leading to speculation that online voting may soon put credentialed experts out of business[2,3]. Recent applications include political and economic forecasting[4,5], evaluating nuclear safety[6], public policy[7], the quality of chemical probes[8], and possible responses to a restless volcano[9]. Algorithms for extracting wisdom from the crowd are typically based on a democratic voting procedure. They are simple to apply and preserve the independence of personal judgment[10]. However, democratic methods have serious limitations. They are biased for shallow, lowest common denominator information, at the expense of novel or specialized knowledge that is not widely shared[11,12]. Adjustments based on measuring confidence do not solve this problem reliably[13]. Here we propose the following alternative to a democratic vote: select the answer that is more popular than people predict. We show that this principle yields the best answer under reasonable assumptions about voter behaviour, while the standard 'most popular' or 'most confident' principles fail under exactly those same assumptions. Like traditional voting, the principle accepts unique problems, such as panel decisions about scientific or artistic merit, and legal or historical disputes. The potential application domain is thus broader than that covered by machine learning and psychometric methods, which require data across multiple questions[14–20].

To illustrate our solution, imagine that you have no knowledge of US geography and are confronted with questions such as: Philadelphia is the capital of Pennsylvania, yes or no? And, Columbia is the capital of South Carolina, yes or no?

You pose them to many people, hoping that majority opinion will be correct. This works for the Columbia question (question C), but most people endorse the incorrect answer (yes) for the Philadelphia question (question P), as shown by the data in Fig. 1a, b. Most respondents may only recall that Philadelphia is a large, historically significant city in Pennsylvania, and conclude that it is the capital[21]. The minority who vote no probably possess an additional piece of evidence, that the capital is Harrisburg. A large panel will surely include such individuals. The failure of majority opinion cannot be blamed on an uninformed panel or flawed reasoning, but represents a defect in the voting method itself.

A standard response to this problem is to weight votes by confidence. For binary questions, confidence c implies a subjective probability c that a respondent's vote is correct and $1-c$ that it is incorrect. Probabilities may be averaged linearly or nonlinearly, producing confidence-weighted voting algorithms[22]. However, these succeed only if correct votes are accompanied by sufficiently greater confidence, which is neither the case for (P) or (C), nor more generally[23]. As shown by Fig. 1c, d, confidences associated with yes and no votes are roughly similar and do not override the incorrect majority in (P).

Here we propose an alternative algorithm that asks respondents to predict the distribution of other people's answers to the question and selects the answer that gains more support than predicted. The intuition underlying the algorithm is as follows. Imagine that there are two possible worlds, the actual one in which Philadelphia is not the capital of Pennsylvania, and the counterfactual one in which Philadelphia is the capital. It is plausible that in the actual world fewer people will vote yes than in the counterfactual world. This can be formalized by the toss of a biased coin where, say, the coin comes up yes 60% of the time in the actual world and 90% of the time in the counterfactual world. Majority opinion favours yes in both worlds. People know these coin biases but they do not know which world is actual. Consequently, their predicted frequency of yes votes will be between 60% and 90%. However, the actual frequency of yes votes will converge to 60% and no will be the surprisingly popular, and correct, answer.

We refer to this selection principle as the 'surprisingly popular' (SP) algorithm, and define it rigorously in the Supplementary Information. In problem (P), the data show that respondents voting yes believe that almost everyone will agree with them, while respondents voting no expect to be in the minority (Fig. 1e). The average predicted percentage of yes votes is high, causing the actual percentage for yes to underperform relative to these predictions. Therefore the surprisingly popular answer is no, which is correct. In (C), by contrast, predictions of yes votes fall short of actual yes votes. The surprisingly popular answer agrees with the popular answer, and the majority verdict is correct (Fig. 1f).

Could an equally valid algorithm be constructed using respondents' confidences? Assume that respondents know the prior world probabilities and coin biases. Each respondent observes the result of their private coin toss, and computes their confidence by applying Bayes' rule. The hypothesized algorithm would need to identify the actual coin from a large sample of reported confidences. Figure 2 proves by counterexample that no such algorithm exists (Theorem 1 in Supplementary Information provides a general impossibility result). It shows how identical distributions of confidences can arise for two different biased coin problems, one where the correct answer is yes and one where the correct answer is no. Admittedly, real people may not conform to the idealized Bayesian model. Our point is that if methods based on posterior probabilities (votes and confidences) fail for ideal respondents, they are likely to fail for real respondents.

By comparison, the SP algorithm has a theoretical guarantee, that it always selects the best answer in light of available evidence (Theorem 2 in Supplementary Information). Theorem 3 extends the algorithm to multiple-choice questions, and shows how vote predictions can identify respondents that place highest probability on the correct answer. These results are based on a common theoretical model that generalizes the biased coin example to multiple, many-sided coins.

To test the SP algorithm, we conducted studies with four types of semantic and perceptual content (details in SI). Studies 1a, b, c used 50 US state capitals questions, repeating the format (P) with different populations. Study 2 employed 80 general knowledge questions.

[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Department of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Princeton Neuroscience Institute and Computer Science Department, Princeton University, Princeton, New Jersey 08544, USA.

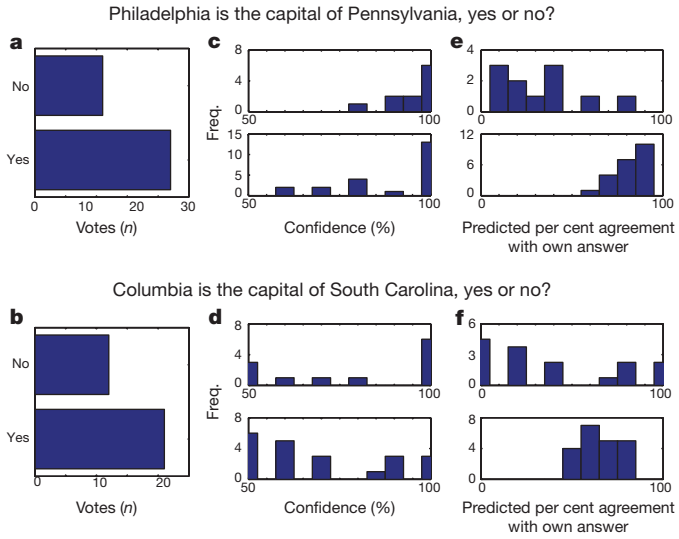


Figure 1 | Two example questions from Study 1c, described in text. **a**, Majority opinion is incorrect for question (P). **b**, Majority opinion is correct for question (C). **c**, **d**, Respondents give their confidence that their answer is correct from 50% (chance) to 100% (certainty). Weighting votes by confidence does not change majority opinion, since respondents voting for both answers are roughly equally confident. **e**, Respondents predict the frequency of yes votes, shown as estimated per cent agreement with their own answer. Those answering yes believe that most others will agree with them, while those answering no believe that most others will disagree. The surprisingly popular answer discounts the more predictable votes, reversing the incorrect majority verdict in (P). **f**, The predictions are roughly symmetric, and so the surprisingly popular answer does not overturn the correct majority verdict in (C).

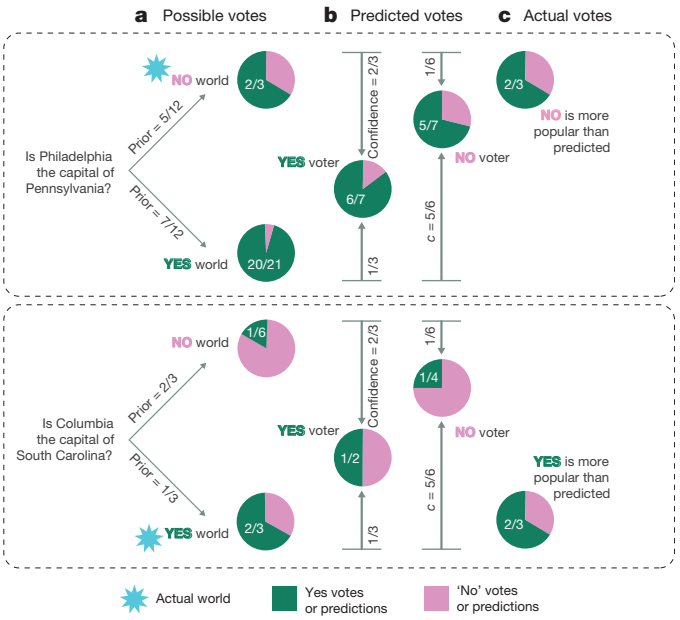


Figure 2 | Why 'surprisingly popular' answers should be correct, illustrated by simple models of Philadelphia and Columbia questions with Bayesian respondents. **a**, The correct answer is more popular in the actual world than in the counterfactual world. **b**, Respondents' vote predictions interpolate between the two possible worlds. In both models, interpolation is illustrated by a Bayesian voter with 2/3 confidence in yes and a voter with 5/6 confidence in no. All predictions lie between actual and counterfactual percentages. The prediction of the yes voter is closer to the percentage in the yes world, and the prediction of the no voter is closer to the percentage in the no world. **c**, Actual votes. The correct answer is the one that is more popular in the actual world than predicted—the surprisingly popular answer. For the Philadelphia question, yes is less popular than predicted, so no is correct. For the Columbia question yes is more popular than predicted, so yes is correct. The example also proves that any algorithm based on votes and confidences can fail even with ideal Bayesian respondents. The two questions have different correct answers, while the actual vote splits and confidences are the same. Confidences 2/3 and 5/6 follow from Bayes' rule if the actual world is drawn according to prior probabilities that favour yes by 7:5 odds on Philadelphia, and favour no by 2:1 odds on Columbia. The prior represents evidence that is common knowledge among all respondents. A respondent's vote is generated by tossing the coin corresponding to the actual world. A respondent uses their vote as private evidence to update the prior into posterior probabilities via Bayes' rule. For example, a yes voter for Philadelphia would compute posterior probability, that is, confidence of $\frac{2}{3}=\frac{7}{12}\times\frac{20}{21}\div\left(\frac{7}{12}\times\frac{20}{21}+\frac{5}{12}\times\frac{2}{3}\right)$ that yes is correct, which is the same confidence computed by a yes voter for Columbia: $\frac{2}{3}=\frac{1}{3}\times\frac{2}{3}\div\left(\frac{1}{3}\times\frac{2}{3}+\frac{2}{3}\times\frac{1}{6}\right)$.

Study 3 asked professional dermatologists to diagnose 80 skin lesion images as benign or malignant. Studies 4a, b presented 90 20th century artworks (Fig. 3) to laypeople and art professionals, and asked them to predict the correct market price category. All studies included a dichotomous voting question, yielding 490 items in total. Studies 1c, 2, and 3 additionally measured confidence. Predicted vote frequencies were computed by averaging all respondents' predictions (details in Supplementary Information).

We first test pairwise accuracies of four algorithms: majority vote, SP, confidence-weighted vote, and max. confidence, which selects the answer endorsed with highest average confidence. Across all 490 items, the SP algorithm reduced errors by 21.3% relative to simple majority vote ($P<0.0005$ by two-sided matched-pair sign test). Across the 290 items on which confidence was measured, the reduction was 35.8% relative to majority vote ($P<0.001$), 24.2% relative to confidence-weighted vote ($P=0.0107$), and 22.2% relative to max. confidence ($P<0.13$).

When frequencies of different correct answers in the same study are imbalanced, percentage agreement can be high by chance. Therefore we assess classification accuracy within a study by categorical correlation coefficients, such as Cohen's kappa, F1 score, or Matthews correlation. The SP algorithm has the highest kappa in every study (Fig. 4); other coefficients yield similar rankings (Extended Data Fig. 1–3).

The art domain, for which majority opinion is too conservative, provides insight into how SP works. Art professionals and laypeople estimated the price of 90 artworks by selecting one of four bins: <$1,000; $1,000–30,000; $30,000–1,000,000; and >$1,000,000. Respondents also predicted the binary division of their sample's votes relative to $30,000. Monetary values throughout refer to US dollars.

Both professionals and laypeople strongly favoured the lower two bins, with professionals better able to discriminate value (Fig. 5). The preference for low price is not necessarily an error. Asked to price an unfamiliar artwork, individuals may rely on their beliefs about market prices, and assume that expensive (>$30,000) pieces are rare. This shared knowledge creates a bias when votes are counted, because similar, hence redundant, base rate information is factored in repeatedly, once for each respondent. Indeed, Fig. 5 shows that the majority verdict is strongly biased against the high category. For example, facing a $100,000 artwork, the average professional has a 30% chance of making the correct call, while the majority vote of the professional panel is directionally correct only 10% of the time. It is difficult for any expensive artwork to be recognized as such by a majority. The SP algorithm corrects this by reducing the threshold of votes required for a high verdict, from 50% to about 25%.

The two studies on propositional knowledge yielded different results (Fig. 4). On capital cities (Studies 1a, b, c), SP reduced the number of incorrect decisions by 48% relative to majority vote. SP was less effective on the knowledge questions in Study 2 (14% error reduction, $P=.031$, two-sided matched-pair sign test). This is the only study that used the Amazon Mechanical Turk respondent pool. In contrast to other studies, the predicted vote splits in Study 2 were in the 40–60% interval for 81%

Figure 3 | Selection of stimuli from Study 4 in which respondents judged the market price of 20th century artworks. a, Roshan Houshmand, *Rhythmic Structure*. **b**, Abraham Dayan, *dance in the living room*. **c**, Matthew Bates, *Botticelli e Filippino*. **d**, Christopher Wool, *Untitled*, 1991, enamel on aluminum, 90″ × 60″ © Christopher Wool; courtesy of the artist and Luhring Augustine, New York. **e**, Anna Jane McIntyre, *Conversation With a Spoonbill*. **f**, Tadeusz Machowski, *Abstract #66*.

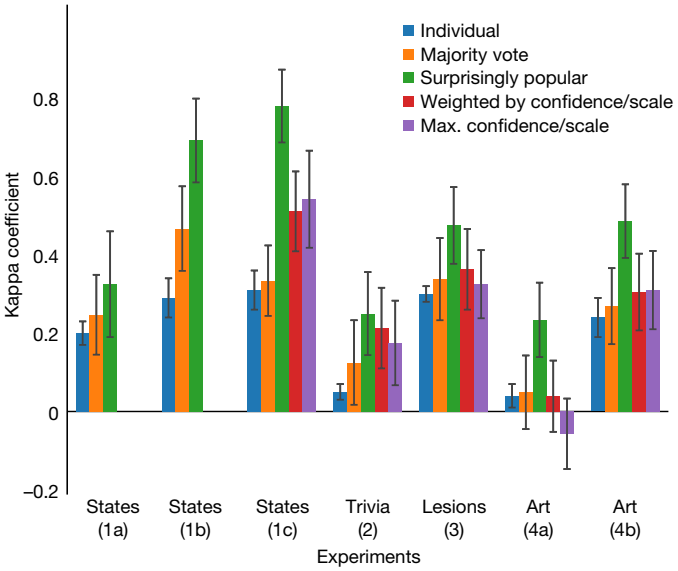


Figure 4 | Results of aggregation algorithms on studies discussed in the text. Study 1a, b, c: n (items per study) = 50; Studies 2 and 3: $n = 80$; Study 4a, b: $n = 90$. Agreement with truth is measured by Cohen's kappa, with error bars showing standard errors. Kappa = $(A - B)/(1 - B)$, where A is per cent correct decisions across items in a study, and B is the probability of a chance correct decision, computed according to answer percentages generated by the algorithm. Confidence was not elicited in Studies 1a, b and 4a, b. However, in 4a, b we use scale values as a proxy for confidence[27], giving extreme categories (on a four-point scale) twice as much weight in scale-weighted voting, and 100% weight in maximum scale. The results for the method labelled 'Individual' are the average kappa across all individuals. SP is consistently the best performer across all studies. Results using Matthews correlation coefficient, F1 score, and per cent correct are similar (Extended Data Figs 1–3).

of items, compared to 22% of such items across other studies. This limited opportunities for SP to alter majority vote.

Empirical results can be compared against simulations based on the biased coin model (Fig. 2). The world prior, coin biases, the actual world, and respondent coin flips are randomly generated to produce simulated finite samples of votes, confidences, and vote predictions (Extended Data Fig. 4 and Supplementary Information). Under these sampling assumptions, individuals are correct 75% of the time. Applying majority voting gives an accuracy of 86%. This 11% improvement is the standard wisdom of the crowd effect. SP is almost infallible for large samples, and it shows good, though not perfect, performance even on small sample sizes. However, given the 86% accuracy of majority vote, SP may need many problems to demonstrate a statistically significant advantage. For example, with 50 problems and $n = 30$, the SP superiority attains $P < 0.05$ for only 40% of simulated studies.

SP performance will always be limited by the information available to the respondents and their competence. If the available evidence is incomplete or misleading, the answer that best fits the evidence may be incorrect. This qualifier can be made explicit by careful phrasing of questions. A question like "Will global temperature increase by more than 5%?" could be worded as: "Given current evidence, is it more likely or not that global temperature will increase by more than 5%?"

The SP algorithm is robust to several plausible deviations from ideal responding (Supplementary Information). The SP outcome will not change, for example, if respondents predict the vote frequency in the world they believe more likely, instead of considering both possible worlds and interpolating predictions (Fig. 2). Alternatively, some respondents may find the prediction task too difficult. In that case, they are likely to predict a 50:50 split or make a random estimate. Such uninformative predictions would move the SP result closer to majority opinion but would not compromise its correct directional influence.

When applying this method to potentially controversial topics, such as political and environmental forecasts, it can be important to guard against manipulation. For example, a respondent might try to increase the chance that a particular option wins by submitting a dishonest low vote prediction for that option. To discourage such behaviour, one can impose truth-telling incentives with the Bayesian truth serum, which also elicits respondents' vote predictions[24,25]. This mechanism scores predictions for accuracy, and answers according to the log-ratio of actual to predicted votes. The log-ratio is an information theoretic measure of surprising popularity, which is maximized by honest responding. Here, we have shown that the surprising popularity of answers is also diagnostic of truth.

The SP algorithm may be compared to prediction markets, where individuals trade contracts linked to specific future events[26]. Both methods allow experts to override the majority view, and both associate expertise with choosing alternatives whose eventual popularity exceeds current expectations. However, unlike prediction markets, SP accepts non-verifiable propositions, such as counterfactual conjectures in public policy, history or law. This, together with the simple input requirements, greatly expands its application range.

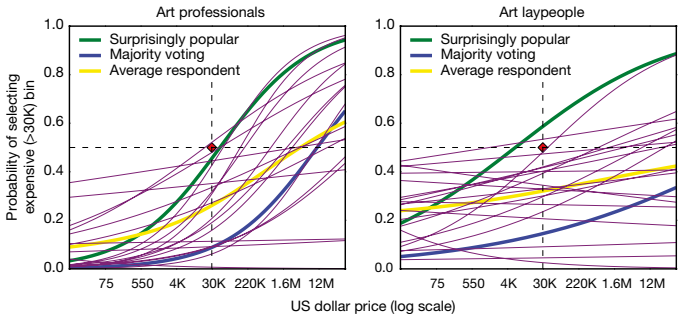


Figure 5 | Logistic regressions showing the probability that an artwork is judged expensive (above $30,000) as function of actual market price. Thin purple lines are individual respondents in the art professionals and laypeople samples, and the yellow line shows the average respondent. Price discrimination is given by the slope of the logistic lines, which is significantly different from zero (χ^2, $P < 0.05$) for 14 of 20 respondents in the professional sample, and 5 of 20 respondents in the laypeople sample (χ^2, $P < 0.05$). Performance is unbiased if a line passes through the red diamond, indicating that an artwork with a true value of exactly $30,000 has a 50:50 chance of being judged above or below $30,000. The bias against the higher price category, which characterizes most individuals, is amplified when votes are aggregated into majority opinion (blue line). The surprisingly popular algorithm (green line) eliminates the bias, and matches the discrimination of the best individuals in each sample.

Although democratic methods of opinion aggregation have been influential and productive, they have underestimated collective intelligence in one respect. People are not limited to stating their actual beliefs; they can also reason about beliefs that would arise under hypothetical scenarios. Such knowledge can be exploited to recover truth even when traditional voting methods fail. If respondents have enough evidence to establish the correct answer, then the surprisingly popular principle will yield that answer; more generally, it will produce the best answer in light of available evidence. These claims are theoretical and do not guarantee success in practice, as actual respondents will fall short of ideal. However, it would be hard to trust a method if it fails with ideal respondents on simple problems like (P). To our knowledge, the method proposed here is the only one that passes this test.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 September; accepted 9 December 2016.

1. Galton, F. Vox populi. *Nature* **75,** 450–451 (1907).
2. Sunstein, C. *Infotopia: How Many Minds Produce Knowledge* (Oxford University Press, USA, 2006).
3. Surowiecki, J. *The Wisdom of Crowds* (Anchor, 2005).
4. Budescu, D. V. & Chen, E. Identifying expertise to extract the wisdom of crowds. Manage. *Sci*. **61,** 267–280 (2014).
5. Mellers, B. *et al.* Psychological strategies for winning a geopolitical forecasting tournament. *Psychol. Sci*. **25,** 1106–1115 (2014).
6. Cooke, R. M. & Goossens, L. L. TU Delft expert judgment data base. *Reliab. Eng. Syst. Saf*. **93,** 657–674 (2008).
7. Morgan, M. G. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc. Natl Acad. Sci. USA* **111,** 7176–7184 (2014).
8. Oprea, T. I. *et al.* A crowdsourcing evaluation of the NIH chemical probes. *Nat. Chem. Biol*. **5,** 441–447 (2009).
9. Aspinall, W. A route to more tractable expert advice. *Nature* **463,** 294–295 (2010).
10. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl Acad. Sci. USA* **108,** 9020–9025 (2011).
11. Chen, K., Fine, L. & Huberman, B. Eliminating public knowledge biases in information-aggregation mechanisms. *Manage. Sci*. **50,** 983–994 (2004).
12. Simmons, J. P., Nelson, L. D., Galak, J. & Frederick, S. Intuitive biases in choice versus estimation: implications for the wisdom of crowds. *J. Consum. Res*. **38,** 1–15 (2011).
13. Hertwig, R. Psychology. Tapping into the wisdom of the crowd–with confidence. *Science* **336,** 303–304 (2012).
14. Batchelder, W. & Romney, A. Test theory without an answer key. *Psychometrika* **53,** 71–92 (1988).
15. Lee, M. D., Steyvers, M., de Young, M. & Miller, B. Inferring expertise in knowledge and prediction ranking tasks. *Top. Cogn. Sci*. **4,** 151–163 (2012).
16. Yi, S. K., Steyvers, M., Lee, M. D. & Dry, M. J. The wisdom of the crowd in combinatorial problems. *Cogn. Sci*. **36,** 452–470 (2012).
17. Lee, M. D. & Danileiko, I. Using cognitive models to combine probability estimates. *Judgm. Decis. Mak*. **9,** 259–273 (2014).
18. Anders, R. & Batchelder, W. H. Cultural consensus theory for multiple consensus truths. *J. Math. Psychol.* **56,** 452–469 (2012).
19. Oravecz, Z., Anders, R. & Batchelder, W. H. Hierarchical Bayesian modeling for test theory without an answer key. *Psychometrika* **80,** 341–364 (2015).
20. Freund, Y. & Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci*. **55,** 119–139 (1997).
21. Goldstein, D. G. & Gigerenzer, G. Models of ecological rationality: the recognition heuristic. *Psychol. Rev*. **109,** 75–90 (2002).
22. Cooke, R. *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford University Press, USA, 1991).
23. Koriat, A. When are two heads better than one and why? *Science* **336,** 360–362 (2012).
24. Prelec, D. A Bayesian truth serum for subjective data. *Science*. **306,** 462–466 (2004).
25. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci*. **23,** 524–532 (2012).
26. Arrow, K. J. *et al.* Economics. The promise of prediction markets. *Science* **320,** 877–878 (2008).
27. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci*. **18,** 1159–1167 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Alam, A. Huang and D. Mijovic-Prelec for help with designing and conducting Study 3, and D. Suh with designing and conducting Study 4b. Supported by NSF SES-0519141, Institute for Advanced Study (Prelec), and Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

Author Contributions All authors contributed extensively to the work presented in this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.P. (dprelec@mit.edu).

Reviewer Information *Nature* thanks A. Baillon, D. Helbing and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Informed consent. All studies were approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). For all studies, informed consent was obtained from respondents using text approved by COUHES. For in-person studies, respondents signed a consent form and for online studies, respondents checked a box.

Studies 1a, b: state capitals. *Materials and methods.* The survey instrument consisted of a single sheet of paper which respondents were asked to complete. The sheet contained 50 propositions each consisting of "X is the capital of Y" for every state Y with X the most populous city in the state Y. For example, the first proposition was "Birmingham is the capital of Alabama". The propositions were in alphabetical order of state. For each proposition, respondents gave the answer T for true or F for false. For each proposition they also estimated the percentage of participants in the experiment who will answer true. There was no time limit for this or any other study.

Respondents and procedure. Study 1a was conducted in the context of two MIT, Sloan MBA classes. A total of 51 respondents were asked to mark their answer sheet by a personal code, and were promised feedback about the results, but no other compensation. Study 1b was conducted at the Princeton Laboratory for Experimental Social Science (PLESS, http://pless.princeton.edu/). Thirty-two respondents were drawn from the pool of pre-registered volunteers in the PLESS database, which is restricted to Princeton students (undergraduate and graduate). Respondents received a flat $15 participation fee. In addition, the two respondents with the most accurate answers received a $15 bonus, as did the two respondents with the most accurate percentage predictions. (In fact, one respondent received both bonuses, earning $45 in total.) Respondents marked their sheet by a pre-assigned code, known only to the PLESS administrator who distributed the fee and bonus.

Study 1c: state capitals. *Materials and methods.* The survey was administered on a computer. On each screen, the header was the sentence "X is the capital of Y" as in studies 1a and 1b. There were then four questions as follows:

(a) Is this more likely [t]rue or [f]alse [Answer t or f]:
(b) What is your estimated probability of being correct (50 to 100 percent):
(c) What percentage of other people do you think thought (a) was true [1 to 100 percent]:
(d) What do you think is the average probability that people answered for (b) [50 to 100 percent]:

In this paper, we do not use the response to question (d).

Respondents and procedure. The study was conducted in the MIT Behavioural Research Laboratory. Thirty-three respondents were recruited from the MIT Brain and Cognitive Sciences Department experimental respondents mailing list, with participation restricted to members of the MIT community. Respondents received a $15 participation fee. In addition, the top 20% of respondents with the most accurate answers with respect to ground truth and the top 20% of respondents with the most accurate predictions about the beliefs of others earned a $25 bonus. Respondents were eligible to receive both bonuses. Both types of bonuses were explained in detail to respondents.

Study 2: general knowledge questions. *Materials and methods.* The survey consisted of 80 trivia questions in the domains of history, language, science, and geography. The survey was administered as an online questionnaire and question order was randomized across respondents. The questions were a subset of the 150 questions from the true/false quizzes in these domains on the quiz site Sporcle (http://www.sporcle.com). Two online pilot experiments (of 70 and 80 questions each) were conducted in which respondents were only asked whether they thought the answer to each question was true or false, that is, respondents were not asked to make predictions about the answers of others. Using the results of the two pilot experiments, 80 questions were selected by matching the questions for percentage correct; for example, a question that 30% of respondents answered correctly was matched with a question that 70% of respondents answered correctly. This resulted in a balanced final survey with respect to the number of questions the majority answered correctly as well as the number of questions for which the correct answer was false. That is, for half of the 80 questions the actual answer was false, and for half the actual answer was true. Of the 40 questions where the actual answer was false, in the pilot 20 were answered incorrectly by the majority, 1 had a tie vote, and 19 were answered correctly by the majority. Of the 40 questions where the actual answer was true, in the pilot 19 were answered incorrectly by the majority, 1 had a tie vote, and 20 were answered correctly by the majority.

Examples of propositions which respondents evaluated, together with the percentage of respondents who answered correctly in the pilot experiment in parentheses, are as follows: Japan has the world's highest life expectancy (10%), Portuguese is the official language of Mozambique (30%), The currency of Switzerland is the Euro (50%), The Iron Age comes after the Bronze Age (70%), The longest bone in the human body is the femur (90%).

Respondents were asked for each question to make their best guess as to whether the proposition is more likely true or false, to think about their own beliefs and estimate the probability that their answer was correct, and to think about other people's beliefs and predict the percentage of people who guessed the answer was true.

To give an estimate of the probability that their answer was correct, respondents chose one of the six following options:

(a) Totally uncertain, a coin toss (about 50% chance of being correct).
(b) A little confident (about 60% chance of being correct).
(c) Somewhat confident (about 70% chance of being correct).
(d) High confidence (about 80% chance of being correct).
(e) Very high confidence (about 90% chance of being correct).
(f) Certain (about 100% chance of being correct).

Respondents were asked not to search for the answers to the questions. Respondents searching for the answer, rather than answering from their own knowledge, does not affect testing the aggregation method since this is simply an additional source of information for some respondents who may thus be more accurate. The average time to complete all three parts of a question was 17 s and it was not the case that if a respondent took more time to answer a question they were more likely to be correct, suggesting that, in fact, searching for the correct answer was not common.

Respondents and procedure. Respondents were recruited from Amazon Mechanical Turk and were paid a flat fee of $5.00 with 39 respondents completing the survey. Respondents who took part in either of the pilot experiments were excluded from participating in the final experiment.

Study 3: dermatologists assessing lesions. *Materials and methods.* The survey was administered online. Respondents were divided into two groups, with one survey containing images of 40 benign and 20 malignant lesions, and the other survey containing images of 20 benign and 40 malignant lesions. The 80 images used in the experiment were obtained from Atlas Dermatologico, DermIS, and DermQuest. The images were selected to be approximately the same size, had no visible signs of biopsy, and were filtered for quality by an expert dermatologist. Question order was randomized across respondents. Since all lesions pictured in the survey had been biopsied, whether a particular lesion was benign or malignant was known to us.

For each image of a lesion, respondents predicted whether the lesion was benign or malignant, gave their confidence on a six point Likert scale from 'absolutely uncertain' to 'absolutely certain' and estimated the likely distribution of opinions amongst other dermatologists on an eleven-point scale from 'perfect agreement that it is benign' to 'perfect agreement that it is malignant' with the midpoint labelled as 'split in opinions with equal number of benign and malignant diagnoses'.

Respondents and procedure. Dermatologists were recruited by referral and 25 respondents answered the survey, with 12 in the condition with 40 benign lesions and 13 in the condition with 20 benign lesions. Respondents had an average of 10.5 years of experience. Respondents were told that a $25 donation would be made to support young investigators in dermatology for every completed survey, and that if the survey was completed by a particular date this would be increased to $50. Respondents were also told that a randomly selected respondent would receive $1,000.

Study 4a, b: professionals and laypeople judging art. *Materials and methods.* The survey instrument consisted of a bound booklet with each page containing a colour picture of a 20th century art piece and questions about the piece. The medium and dimensions were given for each piece.

Respondents were told that the survey contained 90 reproductions of modern (20th century) artworks, and that for each artwork they would be asked a few questions that would help us understand how professionals and non-professionals respond to modern art, including predicting how other people will respond to each piece. Respondents were told that 'professionals' refers to people working with art in galleries or museums, and 'non-professionals' refers to MIT master's and doctoral students who have not taken any formal art or art history classes.

For each artwork, respondents were asked for four pieces of information:

(1) Their 'simple personal response' to the artwork by circling either 'thumbs up' or 'thumbs down'.
(2) Their estimate of the percentage of art professionals and of MIT students circling 'thumbs up' in (1).
(3) Their prediction of the current market price of the artwork by checking one of four value categories: <$1,000; $1,000–30,000; $30,000–1,000,000; and >$1,000,000.

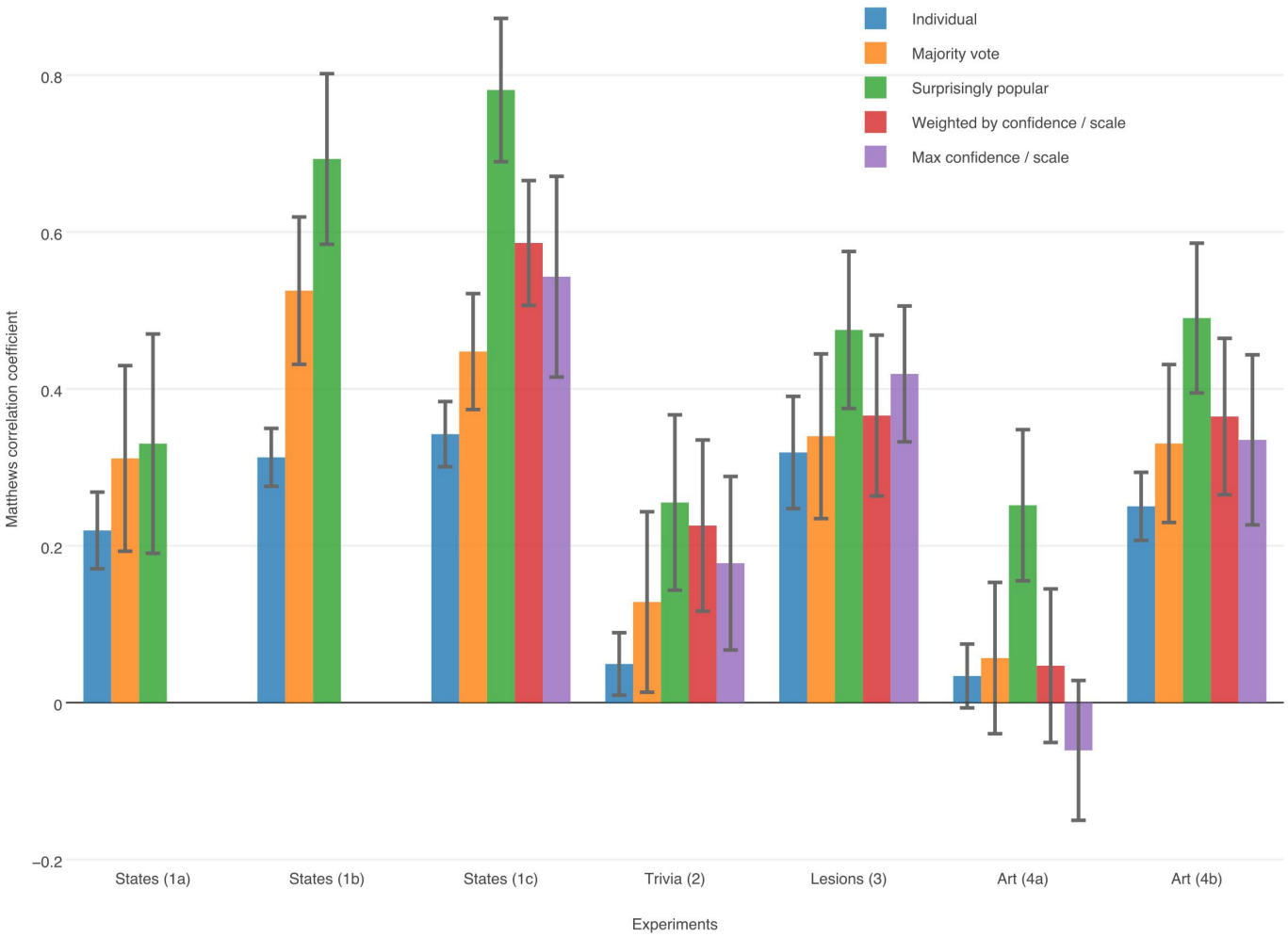
(4) Their estimate of the percentage of art professionals and of MIT students predicting a market value over $30,000.

In this paper, we do not use the responses to questions (1) and (2).

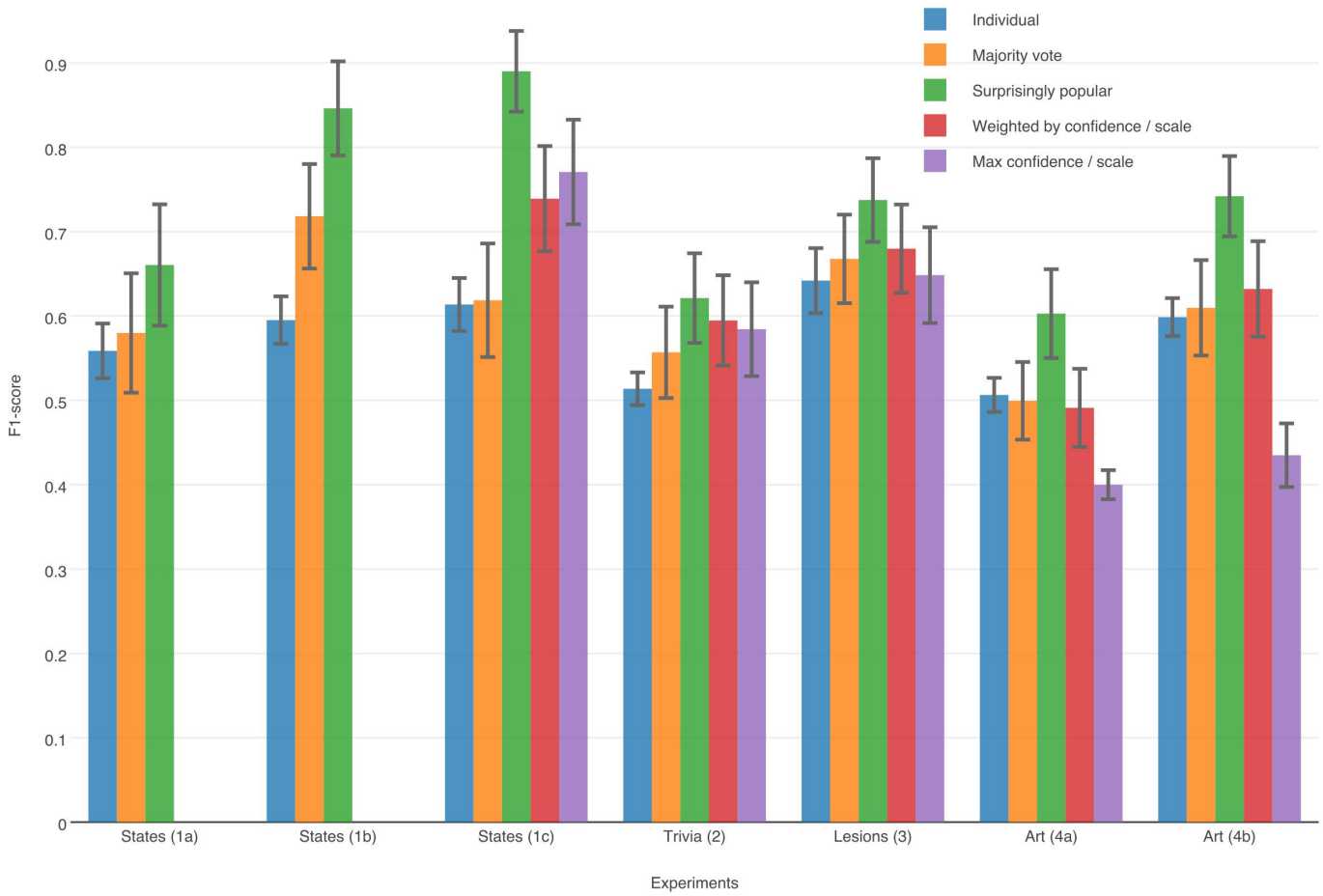
The images in Fig. 4 are reproduced with the permission of the artists and galleries, as indicated in the legend.

Respondents and procedure. Two groups of respondents completed the survey. The MIT group consisted of 20 MIT graduate students who had not taken courses in art or in art history. They were paid $20 as compensation for their time. Respondents came individually into the laboratory, and completed the survey in a room alone. The professional group consisted of art professionals—predominantly managers of art galleries. The art professionals were visited by appointment at their offices and completed the survey during the appointment.

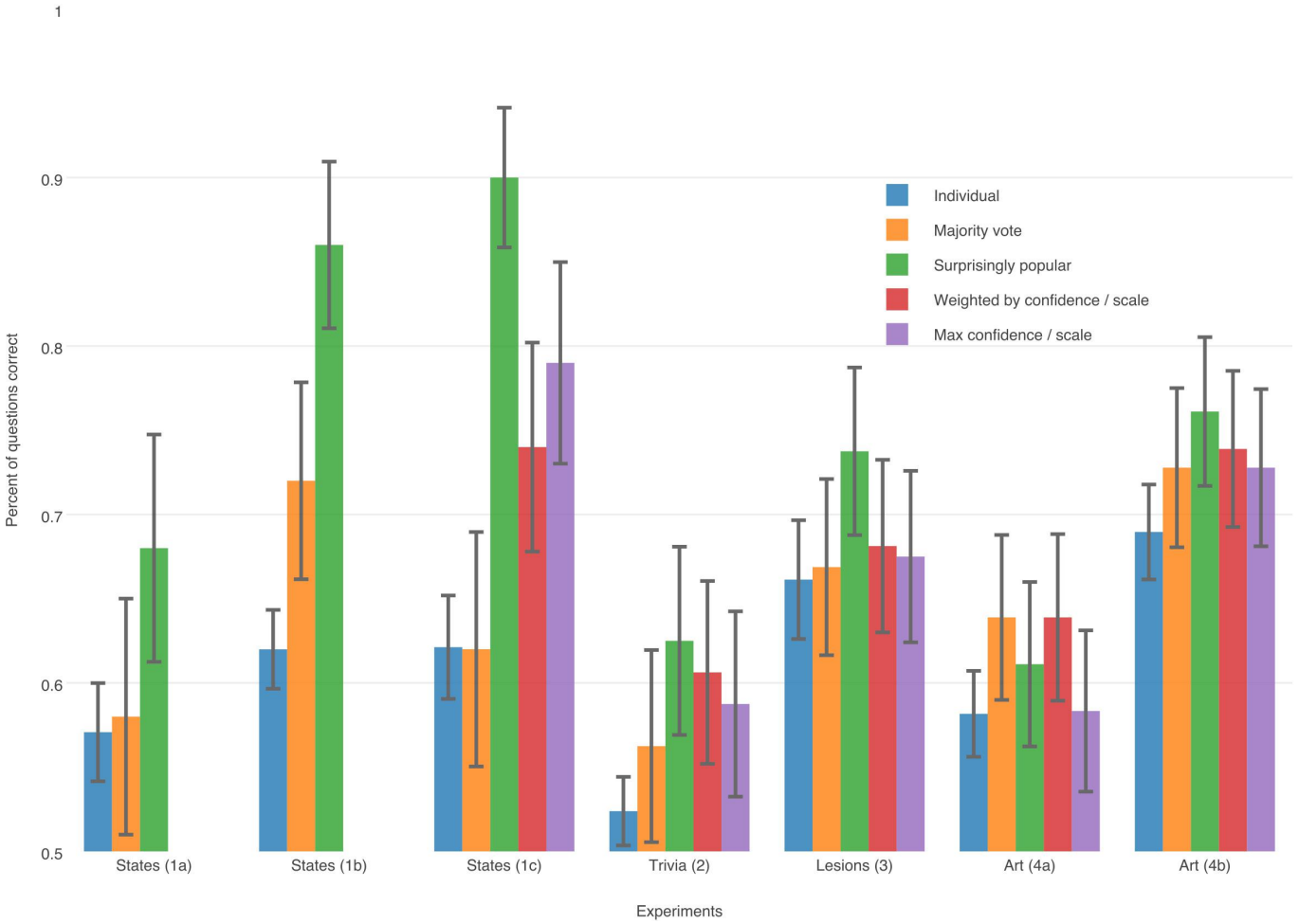
Data availability statement. Data from all studies, as well as analysis code, is available upon reasonable request from the corresponding author.



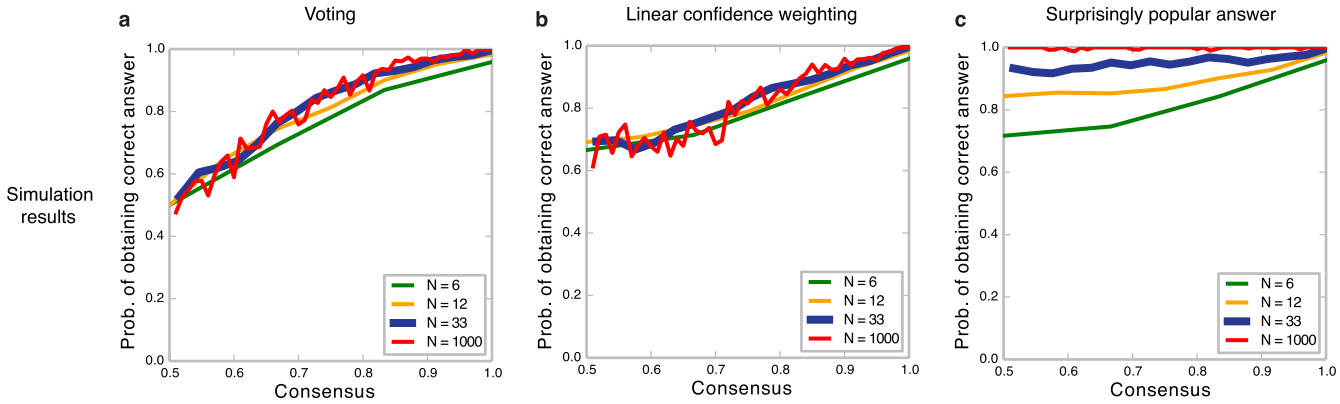
Extended Data Figure 1 | Performance of all methods across all studies, shown with respect to the Matthews correlation coefficient. Error bars are bootstrapped standard errors. Details of studies are given in Fig. 4 of the main text.



Extended Data Figure 2 | Performance of all methods across all studies, shown with respect to the macro-averaged F1 score. Error bars are bootstrapped standard errors. Details of studies are given in Fig. 4 of the main text.



Extended Data Figure 3 | Performance of all methods across all studies, shown with respect to percentage of questions correct. Error bars are bootstrapped standard errors. Details of studies are given in Fig. 4 of the main text.



Extended Data Figure 4 | Performance of aggregation methods on simulated datasets of binary questions, under uniform sampling assumptions. One draws a pair of coin biases (that is, signal distribution parameters), and a prior over worlds, each from independent uniform distributions. Combinations of coin biases and prior that result in recipients of both coin tosses voting for the same answer are discarded. An actual coin is sampled according to the prior, and tossed a finite number of times to produce the votes, confidences, and vote predictions required by different methods (see Supplementary Information for simulation details). As well as showing how sample size affects different aggregation methods the simulations also show that majorities become more reliable as consensus increases. A majority of 90% is correct about 90% of the time, while a majority of 55% is not much better than chance. This is not due to sampling error, but reflects the structure of the model and simulation assumptions. According to the model, an answer with *x*% endorsements is incorrect if counterfactual endorsements for that answer exceed *x*% (Theorem 2), and the chance of sampling such a problem diminishes with *x*.

SUPPLEMENTARY INFORMATION
doi:10.1038/nature21054

A solution to the single question crowd wisdom problem: Supplementary Information

Drazen Prelec, H. Sebastian Seung, and John McCoy
dprelec@mit.edu, sseung@princeton.edu, jmccoy@mit.edu

Contents

1 Theory

The formal model builds on the biased coin example, generalizing to m coins, each coin having n possible sides. In part, this is a standard normative account of how individuals should make inferences about hypotheses (coins) from data (toss outomes). The additional assumption is that individuals take these inferences one step further,

and compute correct expectations of tosses observed by others. We also assume an infinite sample of respondents.

A more complete Bayesian model would have parameters for respondent errors and biases, and would also deal with the finite sample issue. However, it is important to first understand what can be deduced from different types of input in the ideal case. This sets boundaries on what one might expect to achieve with richer models.

We begin with a negative result, Theorem 1, that an infinite sample of correctly computed posterior probabilities over coins is compatible with any possible coin (=answer) being correct. This reveals a limitation of methods based on posterior probabilities, such as votes and confidences.

We then show how to determine the correct answer in three increasingly complex settings: (1) $m = 2, n \geq 2$, (2) $m = n > 2$, (3) $m, n \geq 2$. In particular, Theorem 2 proves that the surprisingly popular algorithm for binary questions described in the main text is valid.

Extensions to multiple choice questions ($m > 2$ coins) rely on a key Lemma, which Theorem 3 applies this to the $m = n > 2$ case. We then consider the fully general m, n problem, and indicate, without formal proof, how the correct answer can be derived if in addition to vote predictions one also elicits posterior probabilities (whose elicitation is not required for Theorems 2 and 3).

The results presented here which justifying choosing the surprisingly popular answer assume ideal Bayesian respondents. However, the biased coin argument presented in the main text remains valid even with certain departures from Bayesian rationality. For example, respondents might simplify the prediction task by predicting the vote split in the world that they think is more likely, and ignoring the possibility of the less likely world. Then, those voting for the correct answer will make accurate predictions, while those voting for the wrong answer will underestimate the vote for the correct answer. The average predicted vote for the correct answer will again underestimate the actual vote, confirming the surprisingly popular principle.

1.1 Model

The model extends the biased coin example in two ways. First, we generalize to an arbitrary number m of possible worlds (each containing a possible coin). One of the worlds is actual, the rest are counterfactual. We identify worlds with possible answers to a multiple choice question. Uncertainty about the actual world, i.e., the correct answer, is modeled by a random variable taking on values in the set $\{a_1, \ldots, a_m\}$ of m possible answers to a multiple-choice question. Second, we distinguish between a respondent's vote for a particular answer and the evidence on which that vote is based. The evidence respondent r possesses is summarized by a private 'signal' S^r, which is a random variable taking on categorical values in the set $\{s_1, \ldots, s_n\}$. A respondent's vote V^r is given by a function $V^r = V(S^r)$ that maps signals to votes $V^r \in \{v_1, \ldots, v_m\}$ for the m possible answers.

Conditional on world a_i, signals of different respondents are independent, identically distributed with probabilities $p(s_k|a_i)$. Therefore, all differences in knowledge are captured by signals. The prior $p(a_i)$ gives probabilities consistent with the evidence that is common knowledge among all respondents. For problem (P) discussed in the main text, common knowledge might be that Philadelphia is a large city. Ideal respondents know the joint distribution $p(s_k, a_i)$, which defines the possible world model (to avoid degeneracies, we assume $p(a_i) > 0, p(s_k) > 0$). However, they do not know which a_i is the correct answer a_{i^*}, and nor do they know the actual distribution of received signals. In terms of the coin example, they know which coins are possible and the properties of each coin, but they do not know which coin is actually being used.

Respondents have two types of beliefs, both computed from their received signal s_k and the joint distribution $p(s_k, a_i)$. Beliefs about the correct answer are given by the posterior probabilities $p(a_i|s_k)$, which can be obtained from knowledge of the joint distribution of signals and answers. Beliefs about signals received by other respondents, say the probability of another respondent receiving signal s_j written as $p(s_j|s_k)$, are derived by computing the distribution of signals $p(s_j|a_i)$ conditional on a particular answer being correct, and marginalizing over all possible answers,

$$p(s_j|s_k) = \sum_i p(s_j|a_i)p(a_i|s_k)p(a_i)$$

More explicitly, one would write, $p(s_j^q|s_k^r) = \Pr(S^q = s_j|S^r = s_k)$, which is the probability that another, randomly selected respondent q receives signal s_j given that respondent r has received signal s_k. We omit the supercripts because the probability is the same for any pair of different respondents q, r.

As discussed in the main text and proven in Theorem 1 below, the probabilities $p(a_i|s_k)$ are always inconclusive, in that even an infinite sample of perfectly computed posterior probabilities over answers is compatible with any given answer being correct in some possible world model. Posterior probabilities strongly constrain the set of models with which they are compatible, but they do not identify the actual world.

Theorem 1. *The correct answer cannot be deduced by any algorithm relying exclusively on knowledge of actual signal probabilities, $p(s_k|a_{i^*}), k = 1, \ldots, n$ and posterior probabilities over answers implied by these signals, $p(a_i|s_k), k = 1, \ldots, n, i = 1, \ldots, m$.*

Proof. The proof is by construction of a possible world model that generates these signal probabilities and posterior probabilities for an arbitrarily selected answer.

Assume that the distribution of signals, $p(s_k|a_{i^*})$, and posterior probabilities, $p(a_j|s_k)$, are known but the correct answer a_{i^*} is unknown. We choose any answer a_i, and construct a corresponding possible world model $q(s_k, a_j)$ such q would generate the known signal distribution and posteriors if $i^* = i$.

Observe first that the known parameters do do not constrain the prior over signals, which we can set equal to:

$$q(s_k) = \frac{p(s_k|a_{i^*})}{p(a_i|s_k)}(\sum_j \frac{p(s_j|a_{i^*})}{p(a_i|s_j)})^{-1},\ k = 1, ..., n$$

Because posteriors must match observed posteriors: $q(a_j|s_k) = p(a_j|s_k)$, for $k = 1, \ldots, n$, $j = 1, \ldots, m$, the possible world model is now fixed: $q(s_k, a_j) = q(a_j|s_k)q(s_k)$. In particular, the prior over answers may be computed from the joint distribution,

$$q(a_i, s_k) = q(a_i|s_k)q(s_k) = p(s_k|a_{i^*})(\sum_j \frac{p(s_j|a_{i^*})}{p(a_i|s_j)})^{-1}$$

by summing over k:

$$q(a_i) = (\sum_j \frac{p(s_j|a_{i^*})}{p(a_i|s_j)})^{-1}$$

The marginal distributions $q(s_k)$, $q(a_i)$, together with the matching posteriors, $q(a_j|s_k) = p(a_j|s_k)$, for $k = 1, \ldots, n$, imply that if the correct answer is a_i, one would observe signal distribution $p(s_k|a_{i^*})$:

$$q(s_k|a_i) = \frac{q(a_i|s_k)q(s_k)}{q(a_i)} = p(s_k|a_{i^*})$$

Because a_i was freely chosen, this proves the theorem. □

Theorem 1 shows that the distribution of posterior probabilities over answers does not rule out any possible answer as the answer responsible for generating that distribution.

We turn therefore to the second type of beliefs, about signals received by other respondents. Because votes are functions of signals, ideal respondents receiving signal s_k can compute the conditional probability $p(v_i|s_k)$ that another respondent will vote for a_i. For example, if the voting function instructs respondents to vote for the most likely answer, $V(s_j) = \text{argmax}_i\, p(a_i|s_j)$, to predict the probability that another respondent votes for a_i the respondent receiving signal s_k would add the probabilities of all signals j that are the most favorable to a_i:

$$p(v_i|s_k) = \sum_{j:V(s_j)=v_i} p(s_j|s_k) = \sum_{i=\text{argmax}_k\, p(a_k|s_j)} p(s_j|s_k)$$

Again, this notation suppresses respondent identity. In explict random variable notation, we would write $p(v_i|s_k)$, as $p(V^q = v_i|S^r = s_k)$ for $q \neq r$, i.e. the probability that an arbitrary respondent q votes for v_i. This is not to be confused with $p(V^r = v_i|S^r = s_k)$ which corresponds to stochastic voting by respondent r.

Similarly, we can define the joint distribution of votes (of an arbitrary respondent) and answers:

$$p(v_i, a_k) = \sum_{j:V(s_j)=v_i} p(s_j, a_j)$$

The conditional distributions, $p(v_i|a_k)$, and $p(a_k|v_i)$, are likewise well defined for any voting function.

1.2 The two worlds, many signals case $m = 2$, $n \geq 2$

We consider a more general version of the voting rule above, which allows us to avoid unanimity even when both signals favor the same answer. Specifically, we consider a cutoff based voting rule that instructs respondents to vote for a_1 if the probability of a_1 exceeds probability c_1, and for a_2 if the probability of a_2 exceeds $c_2 = 1 - c_1$. Formally, we can express this as

$$V(s_k) = \underset{i}{\operatorname{argmax}}\, c_i^{-1} p(a_i|s_k) \tag{1}$$

The above voting rule is identical to the decision algorithm for an ideal observer in signal detection theory. If $c_1 = c_2 = 0.5$, the respondent is assumed to vote for the more likely answer.

Theorem 2. *Assume that not everyone votes for the correct answer. Then the average estimate of the votes for the correct answer will be underestimated.*

Proof. We first show that actual votes for the correct answer exceed counterfactual votes for the correct answer, $p(v_{i^*}|a_{i^*}) > p(v_{i^*}|a_k), k \neq i^*$, as:

$$\frac{p(v_{i^*}|a_{i^*})}{p(v_{i^*}|a_k)} = \frac{p(a_{i^*}|v_{i^*})p(a_k)}{p(a_k|v_{i^*})p(a_{i^*})} = \frac{p(a_{i^*}|v_{i^*})}{(1 - p(a_{i^*}|v_{i^*}))} \frac{(1 - p(a_{i^*}))}{p(a_{i^*})}$$

The fraction on the right is well defined as $0 < p(a_{i^*}|v_{i^*}) < 1$; it is greater than one if and only if $p(a_{i^*}|v_{i^*}) > p(a_{i^*}|v_{i^*})p(v_{i^*}) + p(a_{i^*}|v_k)p(v_k) = p(a_{i^*})$, as $p(a_{i^*}|v_{i^*}) > c_{i^*}$, $p(a_{i^*}|v_k) < c_{i^*}$ by definition of the criterion based voting function.

A respondent with signal s_j computes expected votes by marginalizing across the two possible worlds, $p(v_{i^*}|s_j) = p(v_{i^*}|a_{i^*})p(a_{i^*}|s_j) + p(v_{i^*}|a_k)p(a_k|s_j)$. The actual vote for the correct answer is no less than the counterfactual vote, $p(v_{i^*}|a_{i^*}) \geq p(v_{i^*}|a_k)$. Therefore, $p(v_{i^*}|s_j) \leq p(v_{i^*}|a_{i^*})$, with strict inequality unless $p(a_{i^*}|s_j) = 1$. Because weak inequality holds for all signals, and is strict for some, the average predicted vote will be strictly underestimated. □

If there are more than two possible answers $m > 2$, the actual proportion of votes for the correct answer exceeds predictions provided that votes are defined by a cutoff vector $\sum_i c_i = 1$. However, it no longer points to a unique correct answer, as more than one answer may be underestimated.

1.3 The case $m = n > 2$

Our results for the general case with more than two answers and the same number of signals and answers, rely on a Lemma that shows how the ratio of posterior probabilities on the correct answer relative to any other answer can be derived from the signal frequencies, and their pairwise conditional probabilities. The Lemma is important because it expresses terms whose estimate requires knowing the correct answer (posterior probabilities on truth) as functions of terms that do not require knowledge of the correct answer.

Lemma. *Consider a possible world model with m answers and n signals and joint probability distribution $p(s_j, a_i)$. Let a_{i^*} denote the correct answer. Then:*

$$p(a_{i^*}|s_k) \propto p(s_k|a_{i^*}) \sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)}$$

(setting $0/0 \equiv 0$).

Proof. From Bayes' rule, we have,

$$p(s_i) = p(s_k)\frac{p(s_i|s_k)}{p(s_k|s_i)}$$

After summing over i, with $\sum_i p(s_i) = 1$, we solve for the prior probability of signal s_k:

$$p(s_k) = (\sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)})^{-1}$$

Invoking Bayes' rule again,

$$p(a_{i^*}|s_k) = \frac{p(s_k|a_{i^*})}{p(s_k)}p(a_{i^*}) = p(s_k|a_{i^*}) \sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)}p(a_{i^*})$$

Because $p(a_{i^*})$ is constant across all k, the Lemma follows. □

The Lemma shows how the distribution of signals, and the pairwise predictions of signals, can identify the answer given by respondents who are best informed, in the sense of assigning the highest probability on the correct answer. These respondents would be least surprised by the correct answer, were it revealed.

To convert this Lemma into an algorithm for selecting the correct answer we need to assume that for each answer there is a unique signal such that respondents with that signal assign most probability to this answer, which is also more than the probability assigned to it by other respondents. This assumption is violated, for example, with the posteriors below:

$$p(a_i|s_k) = \begin{pmatrix} .4 & .3 & .3 \\ .45 & .55 & 0 \\ .2 & .3 & .5 \end{pmatrix}$$

If the correct answer is a_1 (first column), then respondents with s_2 (second row) would be least surprised, yet they would believe that the most likely correct answer is a_2. A selection principle based on treating as correct the answer selected by these respondents would incorrectly choose a_2. The theorem below rules out this possibility, by requiring that respondents voting for a given answer assign assign more probability to it than do respondents voting for other answers.

Theorem 3. *Assume $m = n$, $V(s_i) = v_i$, and $p(a_i|s_i) > p(a_i|s_j)$. Let a_{i^*} denote the correct answer. Define the prediction-normalized vote for a_k, $\overline{V}(k)$, as*

$$\overline{V}(k) = p(v_k|a_{i^*}) \sum_i \frac{p(v_i|s_k)}{p(v_k|s_i)}$$

(setting $0/0 \equiv 0$). Then the correct answer has the highest prediction-normalized votes.

Proof. Applying the Lemma, we have,

$$p(a_{i^*}|s_k) \propto p(s_k|a_{i^*}) \sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)}$$

Because $V(s_i) = v_i$, we can rewrite this as:

$$p(a_{i^*}|s_k) \propto p(v_k|a_{i^*}) \sum_i \frac{p(v_i|s_k)}{p(v_k|s_i)} = \overline{V}(k)$$

$p(a_{i^*}|s_{i^*}) > p(a_{i^*}|s_k)$ by the assumption that respondents who vote for a given answer (including a_{i^*}) assign greater posterior probability to it than respondents voting for any other answer a_k. Therefore $\overline{V}(i^*) > \overline{V}(k)$, proving that the correct answer a_{i^*} has the highest prediction-normalized vote. □

One could apply Theorem 3 to experimental data using the following estimation procedure. Because a_{i^*} matches the actual world, the frequency of votes for answer a_k provides an estimate for $p(v_k|a_{i^*})$, which is exact in the limit of an infinite number of respondents. The probabilities $p(v_k|s_i)$ are estimated by asking respondents to predict the frequency of votes v_k and then averaging the predictions of those who voted v_i.

1.4 The case $m, n \geq 2$

It is possible to extend our approach to the general setting of m worlds, and n signals, provided one also elicits respondents' posterior distribution over possible answers.

Here we simply indicate the main idea, and for convenience consider $m > 2$, $n = 2$. That is, we have many coins, each with exactly two sides. Each individual thus has a small amount of information bearing on the many possible answers.

The bias of coin i is given by the ratio on the left side of the Bayesian identity below,

$$\frac{p(s_1|a_i)}{p(s_2|a_i)} = \frac{p(a_i|s_1)}{p(a_i|s_2)} \frac{p(s_1|s_2)}{p(s_2|s_1)}, \; i = 1, ..., n$$

The analyst does not know these true coin biases, but can estimate them from the terms on the right, which respondents provide as their posterior probabilities, $p(a_i|s_1)$, $p(a_i|s_2)$, and pairwise predictions, $p(s_1|s_2)$, $p(s_2|s_1)$. The data therefore can be used to assign the correct bias to each possible coin.

To find the actual coin, the analyst asks respondents to report their toss outcome. The frequencies of observed tosses will converge to $p(s_1|a_{i^*})$ and $p(s_2|a_{i^*})$. The actual coin is then revealed as the coin whose assigned bias matches the observed one:

$$i = i^* \iff \frac{p(a_i|s_1)}{p(a_i|s_2)} \frac{p(s_1|s_2)}{p(s_2|s_1)} = \frac{p(s_1|a_{i^*})}{p(s_2|a_{i^*})}$$

A concrete example illustrates this. Assume three coins, a priori equally likely: (A) 2 : 1 biased for Heads, (B) 2 : 1 biased for Tails, (C) unbiased.

Let us assume the actual coin is C. Respondents report their toss, their posterior probabilities on A, B, C, and their predicted toss distribution. Because toss reports converge to an even split between Heads and Tails, the analyst learns that the actual coin is unbiased. However, he does not yet know which of the coins, A, B, C is the unbiased one. By applying Bayes' rule to their toss, respondents derive and report posterior probabilities over A,B,C as $(\frac{4}{9}, \frac{2}{9}, \frac{1}{3})$ given Heads, and $(\frac{2}{9}, \frac{4}{9}, \frac{1}{3})$ given Tails. With this information, the analyst now knows the exact distribution of posteriors over A, B, C. However, Theorem 1 shows that every such distribution of posteriors can be reconciled with any possible world, that is, any of the three coins.

Adding predictions identifies the actual world. In this case, by symmetry of the assumptions, respondents' predictions are symmetric $p(s_j|s_k) = p(s_j|s_k)$. From the predictions and the posteriors the analyst computes the bias of each possible coin, and notes that coin C is unbiased, and is the only coin whose computed bias matches the actual one. He therefore deduces that the actual coin must be C.

The same method works in the general case, with more than two signals. It is important, however, that the elicitation separates signals (e.g., Heads vs. Tails) and possible states of the world (e.g., A, B, C). Respondents report signals, predict signals, and assign posteriors to states of the world.

1.5 Numerical simulation results

As mentioned in the main text, we performed numerical simulations for the simplest $m = n = 2$ case, under a uniform sampling assumption. We randomly sampled 1000 datasets each with 50 questions, answered by up to 1000 respondents. For each dataset, respondent subsets of size $N \in \{6, 12, 33, 1000\}$ were randomly sampled. For each question $w = 1, 2, \ldots$ a possible world model consisting of a joint distribution of two world states and two signals $p(s_k^w, a_i^w)$ was uniformly sampled, with resampling if it did not satisfy $p(a_k^w|s_k^w) > 0.5$. An actual world a_{i*}^w was sampled given $p(a_i^w)$, and signals were sampled given $p(s_k^w|a_{i*}^w)$. Votes, confidences, and vote predictions were computed for each ideal Bayesian respondent. The results are shown in Extended Data Fig. 4

1.6 Power analysis of the surprisingly popular answer versus voting (based on simulation results)

As described in the section on simulation results above, we sampled 1000 synthetic datasets each consisting of 50 questions and answered by 30 respondents. We computed the voting and surprisingly popular answers and for each dataset performed a paired-samples t-test of voting correctness against the correctness of the surprisingly popular answer. For 392 of the sampled datasets, the test showed a significant advantage ($p < 0.05$) of the surprisingly popular answer over voting, and for none of the sampled datasets was there a significant advantage of voting over the surprisingly popular answer.

2 Experiment Protocols

2.1 Informed consent

All studies were approved by the M.I.T. Committee on the use of humans as experimental subjects (COUHES). For all studies, informed consent was obtained from respondents using text approved by COUHES. For in-person studies, respondents signed a consent form and for online studies, respondents checked a box.

2.2

2.3 Studies 1a, b – State capitals

Materials and methods

The survey instrument consisted of a single sheet of paper which respondents were asked to complete. The sheet contained 50 propositions each consisting of "X is the capital of Y." for every state Y and where X is the most populous city in the state Y.

For example, the first proposition was "Birmingham is the capital of Alabama.". The propositions were in alphabetical order of state. For each proposition, respondents gave the answer T for True or F for False. For each proposition they also estimated the percentage of participants in the experiment who will answer True. There was no time limit.

Respondents and procedure

Study 1a was conducted in the context of two MIT, Sloan MBA classes. 51 respondents were asked to mark their answer sheet by a personal code, and were promised feedback about the results, but no other compensation. Study 1b was conducted at the Princeton Laboratory for Experimental Social Science (PLESS, http://pless.princeton.edu/). 32 respondents were drawn from the pool of pre-registered volunteers in the PLESS database, which is restricted to Princeton students (undergraduate and graduate). Respondents received a flat $15 participation fee. In addition, the two respondents with the most accurate answers received a $15 bonus, as did the two respondents with the most accurate percentage predictions. (In fact, one respondent received both bonuses, earning $45 in total). Respondents marked their sheet by a pre-assigned code, known only to the PLESS administrator who distributed the fee and bonus.

2.4 Study 1c – State capitals

Materials and methods

The survey was administered on a computer. On each screen, the header was the sentence "X is the capital of Y." as in studies 1a and 1b. There were then four questions as follows:

- (a) Is this more likely [t]rue or [f]alse [Answer t or f]:
- (b) What is your estimated probability of being correct (50 to 100 percent):
- (c) What percentage of other people do you think thought (a) was true [1 to 100 percent]:
- (d) What do you think is the average probability that people answered for (b) [50 to 100 percent]:

In this paper, we do not use the response to question (d).

Respondents and procedure

The study was conducted in the MIT Behavioral Research Lab (http://web.mit.edu/brl/). 33 respondents were recruited from the MIT Brain and Cognitive Sciences Department

experimental respondents mailing list, with participation restricted to members of the MIT community. Respondents received a $15 participation fee. In addition, the top 20% of respondents with the most accurate answers with respect to ground truth and the top 20% of respondents with the most accurate predictions about the beliefs of others earned a $25 bonus. Respondents were eligible to receive both bonuses. The explanation given to respondents about the bonus system is reproduced below.

Determination of bonus:

After the study is complete, we will calculate two accuracy scores for each respondent.

(1). Your objective accuracy score is based on your answers to (a) and (b).

For each statement we calculate the probability that you think the statement is true and use this, together with whether the statement is actually true to calculate your score. We use the Brier scoring function, which is designed so that your score is maximized when you report your true guess and confidence level. Below is a table which helps you understand the score you would receive, depending on whether your answer in (a) was correct or incorrect. The table gives the score at intervals of ten percentage points, but you can choose any percentage between 50% and 100%.

Your confidence	*Score if (a) correct*	*Score if (a) incorrect*
50%	*0*	*0*
60%	*9*	*-11*
70%	*16*	*-24*
80%	*21*	*-39*
90%	*24*	*-56*
100%	*25*	*-75*

Points to note:

- *the more certain you claim to be, the more points you can win*
- *as you approach 100%, the penalty for being incorrect climbs much faster than the gains for being correct.*

A tip:

- *In the long run, you will score the most points if the numbers correspond to your true levels of confidence. Expressing too much confidence is a common mistake in this game.*

(2). Your prediction accuracy score is based on your answers to (c) and (d).

Your prediction accuracy score reflects how well you have predicted the actual percentages of respondents who answered Yes to each of the fifty questions, and how well you have estimated the average confidence levels.

2.5 Study 2 – General knowledge questions

Materials and Methods

The survey consisted of 80 trivia questions in the domains of history, language, science, and geography. The survey was administered as an online questionnaire and question order was randomized across respondents. The questions were a subset of the 150 questions from the True/False quizzes in these domains on the quiz site Sporcle (www.sporcle.com). Two online pilot experiments (of 70 and 80 questions each) were conducted in which respondents were only asked whether they thought the answer to each question was True or False, i.e respondents were not asked to make predictions about the answers of others. Using the results of the two pilot experiments, 80 questions were selected by matching the questions for percentage correct, e.g. a question that 30% of respondents answered correctly was matched with a question that 70% of respondents answered correctly. This resulted in a balanced final survey with respect to the number of questions the majority answered correct as well as the number of questions for which the correct answer was false, as shown by the contingency table in Table S1.

	Actual answer is false	Actual answer is true	
Majority incorrect	20	19	39
Majority tie	1	1	2
Majority correct	19	20	39
	40	40	80

Table S1: Contingency table showing distribution of questions for Study 2.

Example questions, together with the percentage of respondents who answered correctly in the pilot experiment are shown in Table S2.

Example question	Percent of respondents correct in pilot experiments
Japan has the world's highest life expectancy	10
The Nile River is more than double the length of the Volga	20
Portuguese is the official language of Mozambique	30
Avogadro's constant is greater than Planck's constant	40
The currency of Switzerland is the Euro	50
Abkhazia is a disputed territory in Georgia.	50
The chemical symbol for Tin is Sn	60
The Iron Age comes after the Bronze Age	70
Schuyler Colfax was Abraham Lincoln's Vice President	80
The longest bone in the human body is the femur	90

Table S2: Example questions from Study 2 and percent correct in pilot experiments.

Respondents were given the following instructions:

Please read the following 80 True/False trivia questions carefully and make your best guess.

For each question, we'll ask you to do three things:

(a) *Say whether you think the statement is more likely True or False*

(b) *Think about your own beliefs and estimate the probability that your answer is correct*

(c) *Think about other people's beliefs and predict the percentage of people who guessed the answer was 'True'*

To give an estimate of the probability that their answer was correct, respondents chose one of the six following options:

(a) Totally uncertain, a coin toss (about 50% chance of being correct)

(b) A little confident (about 60% chance of being correct)

(c) Somewhat confident (about 70% chance of being correct)

(d) High confidence (about 80% chance of being correct)

(e) Very high confidence (about 90% chance of being correct)

(f) Certain (about 100% chance of being correct)

To answer the question about other people's votes, respondents gave a percentage.

Respondents were asked to not search for the answers to the questions. Respondents searching for the answer, rather than answering from their own knowledge, does not make affect testing the aggregation method since this is simply an additional source of information for some respondents who may thus be more accurate. The average time to complete all three parts of a question was 17 seconds and it was not the case that if a respondent took more time to answer a question they were more likely to be correct, suggesting that, in fact, searching for the correct answer was not common.

Respondents and Procedure

Respondents were recruited from Amazon Mechanical Turk and were paid a flat fee of $5.00 with 39 respondents completing the survey. Respondents who took part in either of the pilot experiments were excluded from participating in the final experiment.

2.6 Study 3 – Dermatologists assessing lesions

Materials and Methods

The survey was administered online. Respondents were divided into two groups, with one survey containing images of 40 benign and 20 malignant lesions, and the other survey containing images of 20 benign and 40 malignant lesions. The 80 images used in the experiment were obtained from Atlas Dermatologico, DermIS, and DermQuest. The images were selected to be approximately the same size, had no visible signs of biopsy, and were filtered for quality by an expert dermatologist. Question order was randomized across respondents. Since all lesions pictured in the survey had been biopsied, whether a particular lesion was benign or malignant was known to us.

For each image of a lesion, respondents predicted whether the lesions was benign or malignant, gave their confidence on a six point Likert scale from 'absolutely uncertain' to 'absolutely certain' and estimated the likely distribution of opinions amongst other dermatologists on an eleven point scale from 'perfect agreement that it is benign' to 'perfect agreement that it is malignant' with the midpoint labeled as 'split in opinions with equal number of benign and malignant diagnoses'.

Respondents and procedure

Dermatologists were recruited by referral and 25 respondents answered the survey, with 12 in the condition with 40 benign lesions and 13 in the condition with 20 benign lesions. Respondents had an average of 10.5 years of experience. Respondents were told that a $25 donation would be made to support young investigators in dermatology for every completed survey, and that if the survey was completed by a particular date

this would be increased to $50. Respondents were also told that a randomly selected respondent would receive $1000.

2.7 Study 4a, b – Professionals and laypeople judging art

Materials and Methods

The survey instrument consisted of a bound booklet with each page containing a color picture of a 20th century art piece and questions about the piece. The medium and dimensions were given for each piece. Respondents were given the following generalinstructions about the survey:

> *The survey contains 90 reproductions of modern (20th century) artworks. For each artwork we will aks you a few questions.*
>
> *Your answers will help us understand how professionals and non-professionals respond to modern art.*
>
> - *By professionals, we have in mind people working with art, in galleries or museums.*
> - *By non-professionals, we are referring to MIT master's and doctoral students who have not taken any formal art or art history classes.*
>
> *We are also interested in how well people can predict the responses of other people. So, some questions will ask you to guess how other people will respond.*
>
> *This will be explained more fully on the next page. If there is anything unclear about our instructions please do not hesitate to ask!*

For each artwork, respondents were asked for four pieces of information:

- (1) Their 'simple personal response' to the artwork by circling either 'thumbs up' or 'thumbs down'.
- (2) Their estimate of the percentage of art professionals and of MIT students circling 'thumbs up' in (1).
- (3) Their prediction of the current market price of the artwork by checking one of four value categories: under $1,000, or $1,000 to $30,000, or $30,000 to $1,000,00, or over $1,000,000.
- (4) Their estimate of the percentage of art professionals and of MIT students predicting a market value over $30,000.

In this paper, we do not use the responses to questions (1) and (2).

Respondents and Procedure

Two groups of respondents completed the survey. The MIT group consisted of twenty MIT graduate students who had not taken courses in art or in art history. They were paid $20 as compensation for their time. Respondents came individually into the lab, and completed the survey in a room alone. The Newbury group, named for Newbury street in Boston which has many art galleries, consisted of art professionals – predominantly managers of art galleries. The art professionals were visited by appointment at their offices and completed the survey during the appointment.

3 Vote predictions

Our model describes how Bayesian respondents formulate a prediction of the vote distribution of others based on their received signal. Here, we describe some descriptive statistics of the predictions of votes. Respondents' votes were, in general, correlated with their own answers, which is one reason that vote predictions tend not to be simply 50%-50% . That is, respondents voting for option A, compared to those voting for option B, put higher probability on other respondents also voting for option A. Across the three states studies, respondents voting for False predicted that 49% of respondents would endorse True, whereas those voting for True predicted that 73% would. For the trivia study, this prediction was 45% and 59%, respectively. In the lesion study, respondents who voted for benign predicted that 34% of respondents would vote for malignant, but those voting for malignant predicted 79%. In the Art MIT study, those voting for the expensive price bin predicted that 46% of other would, whereas those voting for cheap predicted that only 22% would vote for expensive. For the study with art professionals, these predictions were 53% and 14%, respectively.

For each study, we can examine how close predictions were to within uniform. For studies where predictions were given as percentages, we count the fraction of times that a vote prediction is given which is within 10% of 50-50, i.e. the prediction of votes for the first option is between 40% and 60%. Across the three states studies, an average of 36% of predictions were within 10% of 50-50, in the trivia study 56%, in the art MIT study 29%, and in the art professionals study 19%. For the lesions study, respondents gave their predictions on an 11 point scale, and 30% of predictions were one of the three middle bins.

As a further description of the predictions that people made, we compare predictions of the perecentage of people voting True (or malignant) to the probability that people put on the answer being True (or malignant), inferred from their cote and confidence. For the MIT states study where confidence was elected this correlation is $r_S = 0.64$, $(p < 0.001)$, for the lesions study $r_s = 0.87$, $(p < 0.001)$, and for the trivia study $r_S = 0.48$, $(p < 0.001)$.

We do not have sufficient experimental evidence to justify a particular method

of eliciting useful vote predictions, but we offer a few suggestions, and speculations for future testing. Vote predictions can be incentivized for accuracy. For example, respondents can be paid a bonus which depends on the Kullback-Leibler divergence between their prediction and the actual distribution, or incentivized more generally using the Bayesian Truth Serum. Respondents can be explicitly encouraged to consider whether they are in the minority or majority, and what opinions people different to themselves may hold. Instructions may help respondents recognize cases where despite them having high confidence in an answer, they should also believe that only a minority of respondents would vote for this answer. It is possible that choosing to not elicit confidence prior to eliciting predictions may help respondents to avoid conflating these two quantities. When dealing with respondents answering multiple questions who give identical vote predictions for every question, one could take steps to encourage them to reflect on whether this is an accurate reflection of their beliefs. Respondents could be given information about the composition of the sample that they are in, to aid them in making good predictions about the answers of others.

Note that if all respondents simply predict that 50% of the sample will endorse each of two possible answers, then the surprisingly popular answer is the same as that obtained by majority rule.