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.

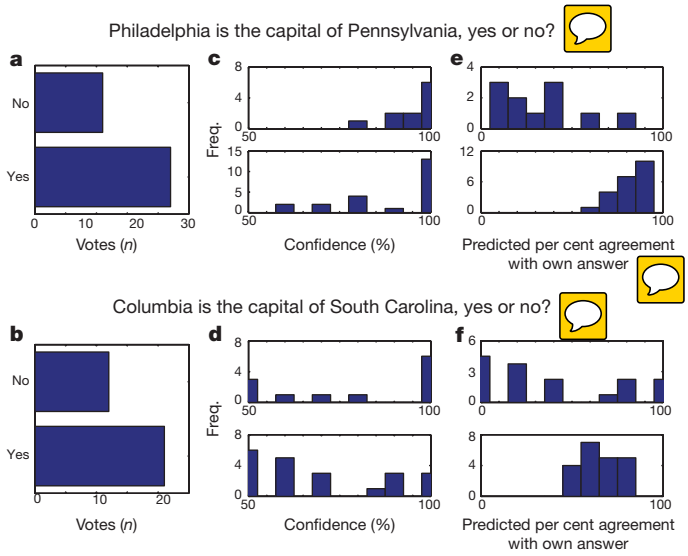


Figure 1 | Two example questions from Study 1c, described in text. **a**, Majority opinion is incorrect for question (P). **b**, Majority opinion is correct for question (C). **c**, **d**, Respondents give their confidence that their answer is correct from 50% (chance) to 100% (certainty). Weighting 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).

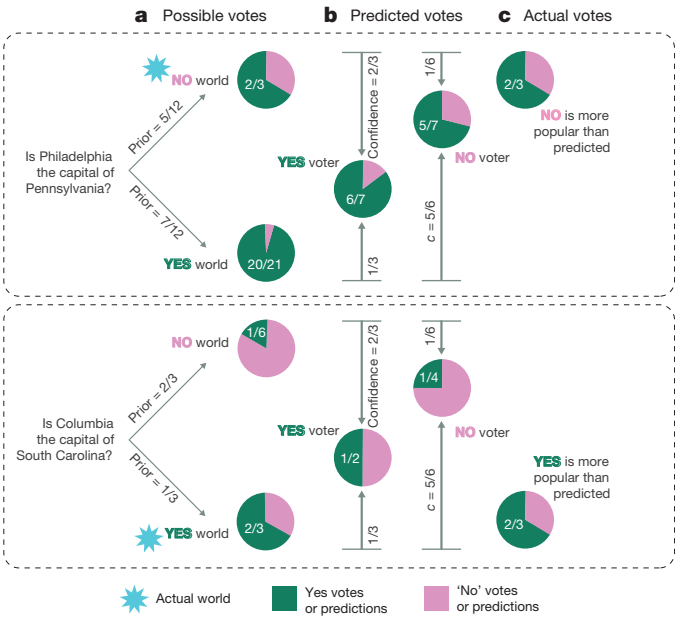


Figure 2 | Why 'surprisingly popular' answers should be correct, illustrated by simple models of Philadelphia and Columbia questions with Bayesian respondents. **a**, The correct answer is more popular in the actual world than in the counterfactual world. **b**, Respondents' vote predictions interpolate 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*.

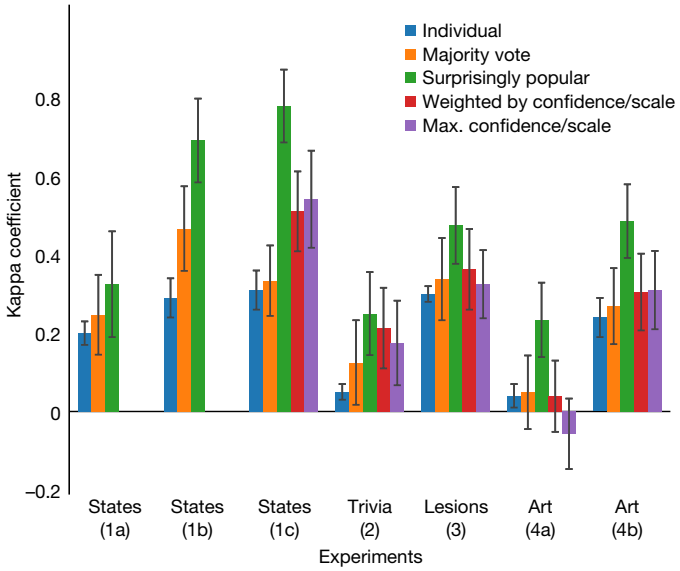


Figure 4 | Results of aggregation algorithms on studies discussed in the text. Study 1a, b, c: n (items per study) = 50; Studies 2 and 3: $n = 80$; Study 4a, b: $n = 90$. Agreement with truth is measured by Cohen's kappa, with error bars showing standard errors. Kappa = $(A - B)/(1 - B)$, where A is 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.

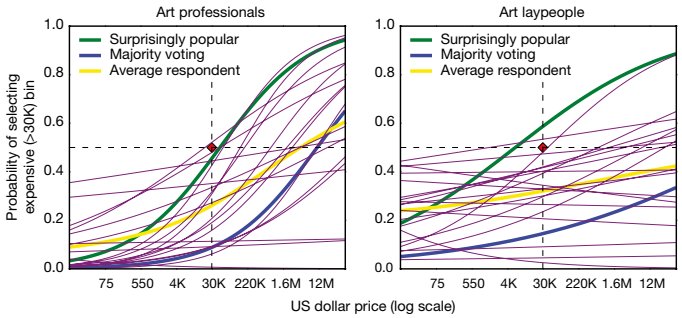


Figure 5 | Logistic regressions showing the probability that an artwork is judged expensive (above $30,000) as function of actual market price. Thin purple lines are individual respondents in the art professionals and laypeople samples, and the yellow line shows the average respondent. Price discrimination 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 le[illegible]. *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.

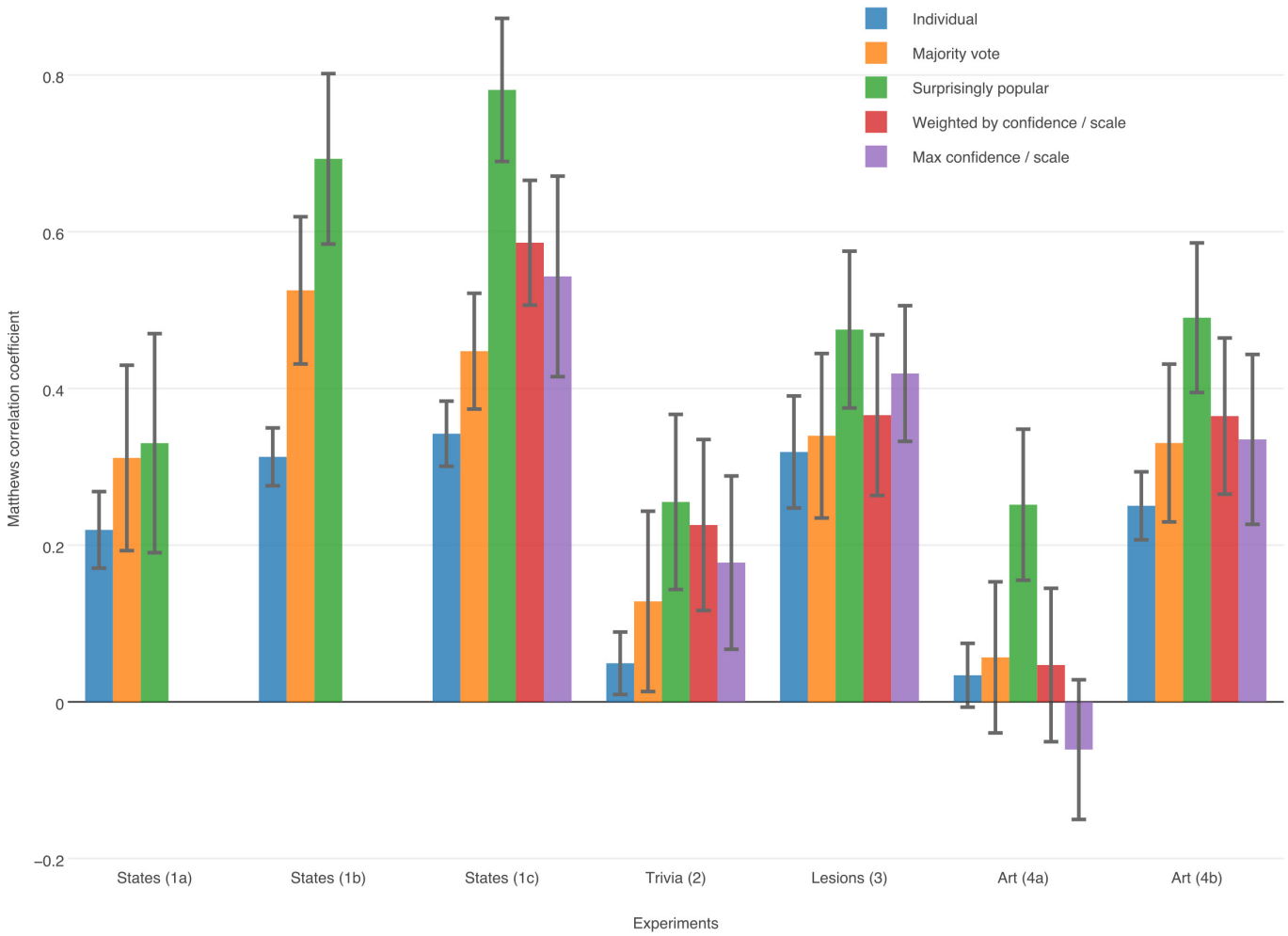
(4) Their estimate of the percentage of art professionals and of MIT students predicting a market value over $30,000.

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

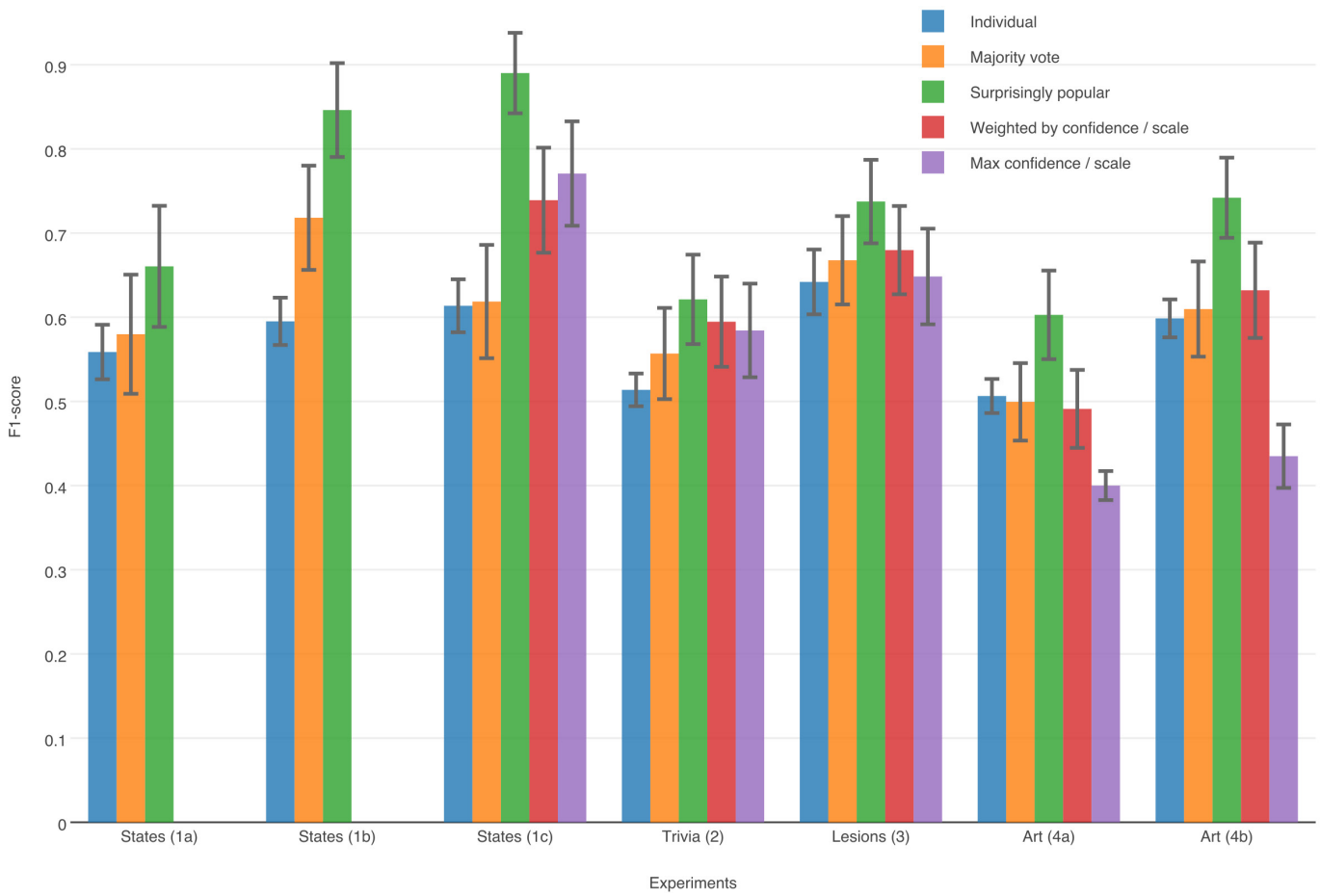
The images in Fig. 4 are reproduced with the permission of the artists and galleries, as indicated in the legend.

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.

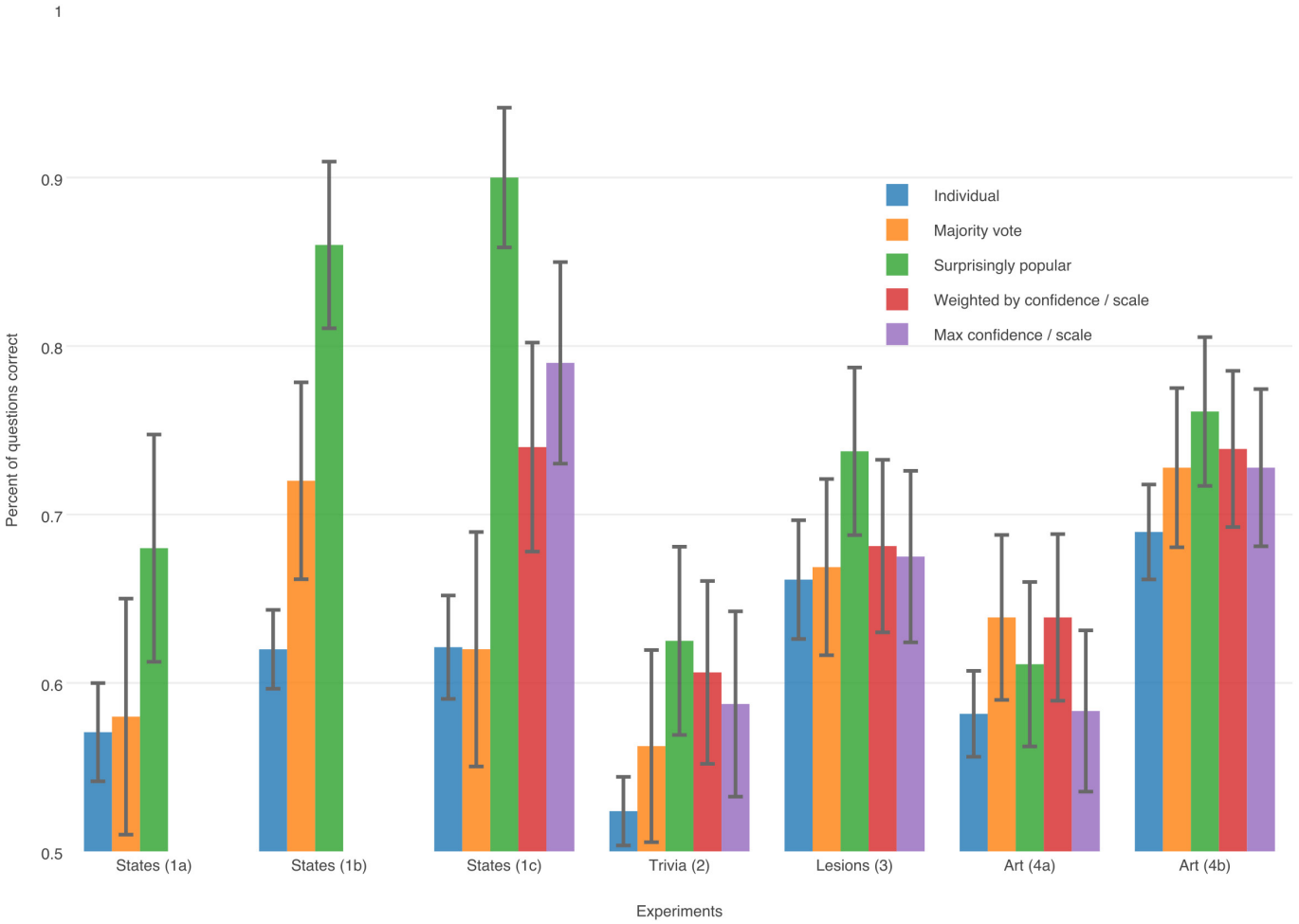
Data availability statement. Data from all studies, as well as analysis code, is available upon reasonable request from the corresponding author.



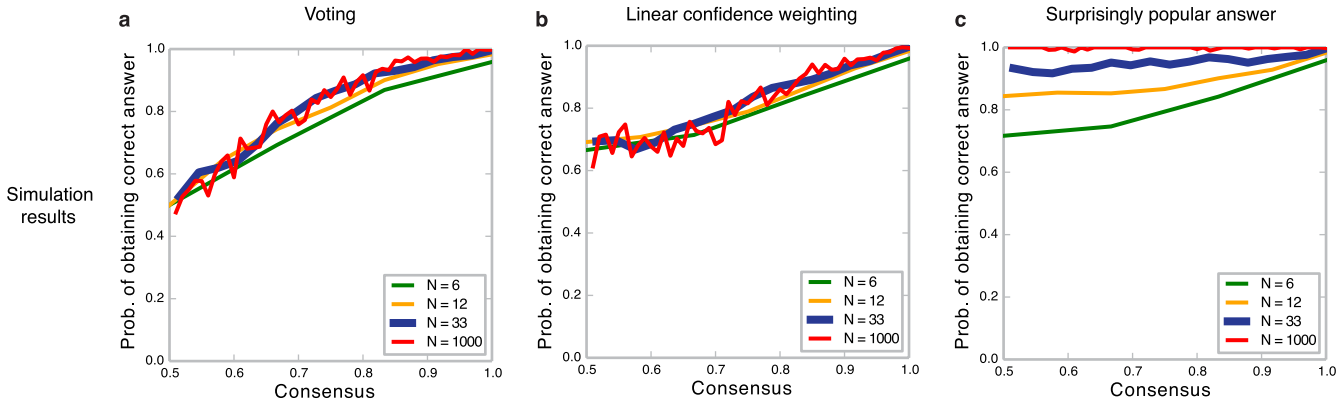
Extended Data Figure 1 | Performance of all methods across all studies, shown with respect to the Matthews correlation coefficient. Error bars are bootstrapped standard errors. Details of studies are given in Fig. 4 of the main text.



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*.