RESEARCH METHODS

How to detect high-performing individuals and groups: Decision similarity predicts accuracy

R. H. J. M. Kurvers[1,2]*, S. M. Herzog[1], R. Hertwig[1], J. Krause[2], M. Moussaid[1], G. Argenziano[3], I. Zalaudek[4], P. A. Carney[5], M. Wolf[2]

Distinguishing between high- and low-performing individuals and groups is of prime importance in a wide range of high-stakes contexts. While this is straightforward when accurate records of past performance exist, these records are unavailable in most real-world contexts. Focusing on the class of binary decision problems, we use a combined theoretical and empirical approach to develop and test a approach to this important problem. First, we use a general mathematical argument and numerical simulations to show that the similarity of an individual's decisions to others is a powerful predictor of that individual's decision accuracy. Second, testing this prediction with several large datasets on breast and skin cancer diagnostics, geopolitical forecasting, and a general knowledge task, we find that decision similarity robustly permits the identification of high-performing individuals and groups. Our findings offer a simple, yet broadly applicable, heuristic for improving real-world decision-making systems.

INTRODUCTION

Identifying high-performing individuals and collectives is a key challenge in a wide range of high-stakes contexts, from medical and psychological diagnostics and lie detection to economic and political forecasting, environmental risk analyses, and investment decisions (*1*–*9*). Telling them apart is relatively straightforward when accurate records of past performance exist (*5*, *9*, *10*). In many real-world contexts, however, these records do not exist or are inaccessible due to ethical or legal constraints (*11*–*13*). The most common approach, then, is to identify high performers based on proxies that are thought to correlate with decision accuracy, such as experience or reputation, or self-identified or peer-assessed expertise (*14*–*16*). The purported correlations between accuracy and these proxies, however, are often poorly understood, complex, and context dependent (*4*, *8*, *16*). Focusing on the class of binary decision problems, we here use a combined theoretical and empirical approach to develop and test a previously unidentified method for selecting high-performing individuals and groups—crucially, this method relies solely on inputs that are readily observable and does not require information about past decision accuracy of individuals.

We proceed as follows. First, we use a general mathematical argument to show that—in any binary decision task in which individuals are, on average, more often correct than not—the similarity of an individual's decision to those of others (i.e., the average percentage agreement with others) is tightly correlated with that individual's decision accuracy. Second, using numerical simulations, we show that this correlation between decision similarity and accuracy is observed even when one relaxes the two simplifying assumptions underlying our analytical result (i.e., when decisions of different individuals are correlated and/or when decision similarity is calculated from small samples)—as long as the average accuracy of decision makers exceeds 0.5. Third, using several large datasets from three different domains—medical diagnostics, geopolitical forecasting, and general knowledge—we show that (i) in each of these domains, as predicted, the decision similarity of an individual is tightly correlated with that individual's decision accuracy, and (ii) this association can be exploited to reliably detect both high-performing individuals and groups.

RESULTS

Analytical result: Decision similarity correlates with decision accuracy

Consider a pool of N decision makers facing a binary decision problem (e.g., dermatologists classifying a skin lesion as benign or malignant; forecasters predicting whether a regime will still be in power 1 year from now), individuals differ in their average accuracy p_i, $i = 1 \ldots N$. Suppose we confront each individual with the same set of cases (e.g., skin lesions, forecasts), and let us compare the decisions of each individual to the decisions of a "benchmark individual" judging the same cases. More specifically, for each individual i, we define the decision similarity S_i as the fraction of cases for which this individual makes the same decision as the benchmark individual. The expected value of this decision similarity is

$$E(S_i) = p_i \cdot p + (1 - p_i) \cdot (1 - p) \quad (1)$$

where p is the probability that the benchmark individual is correct in any particular case. To see how our similarity measure depends on individual accuracy levels, we compare the expected decision similarity (to the benchmark individual) for two individuals j and i, which is given by

$$E(S_j) - E(S_i) = (p_j - p_i) \cdot (2 \cdot p - 1) \quad (2)$$

Crucially, as can be seen from Eq. 2, whenever the benchmark individual is more often correct than incorrect (i.e., $p > 0.5$), individuals with a higher accuracy will be characterized by a higher expected decision similarity, that is, $p_j > p_i \Leftrightarrow E(S_j) > E(S_i)$. That is, the higher an individual's decision similarity, the more accurate this individual's decisions are, provided the benchmark individual performs above

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. [2]Leibniz Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 310, 12587 Berlin, Germany. [3]Dermatology Unit, University of Campania, Naples 80131, Italy. [4]Dermatology Clinic, Maggiore Hospital, University of Trieste, Piazza dell' Ospedale 1, 34125 Trieste, Italy. [5]Department of Family Medicine, Knight Cancer Institute, Oregon Health & Science University, Portland, OR 97239, USA.
*Corresponding author. Email: kurvers@mpib-berlin.mpg.de

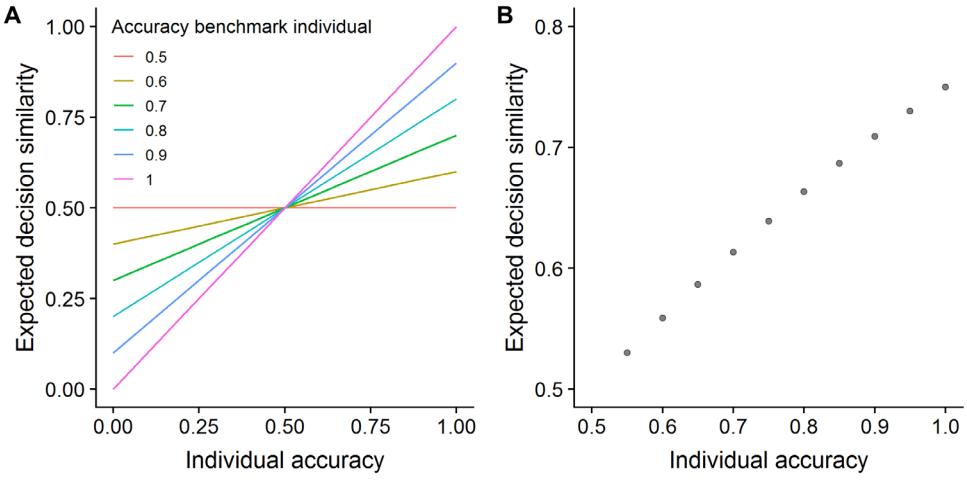

Fig. 1. Analytical prediction: Decision similarity is tightly associated with decision accuracy in binary decision problems. (**A**) When compared with a benchmark individual with $p > 0.5$, the expected decision similarity $E(S_i)$ of individual i increases with its accuracy level p_i; the different lines correspond to benchmark individuals with different levels of accuracy. (**B**) When comparing the decisions of individuals to the decisions of all other individuals in a pool of candidate decision makers, the expected average decision similarity of individuals is tightly correlated with individual accuracy, as long as the average accuracy of individuals in the pool is above 0.5 (more precisely, as long as the average accuracy of individuals remains above 0.5 after excluding every possible pair of individuals, see Eq. 4). The panel illustrates this for a pool of 10 decision makers with decision accuracies of 0.55, 0.60, 0.65,…1.0, respectively. The expected decision similarity is calculated using Eq. 3.

chance level. Figure 1A illustrates this tight relationship between decision similarity and accuracy when comparing a focal individual to different benchmark individuals.

To illustrate the mechanism underlying the similarity-accuracy relationship as clear as possible, we have, up to now, considered the case where we compare the decisions of individuals to the decisions of a single benchmark individual. In practice, however, one may not know whether a particular decision maker performs above chance level. This problem can be solved by comparing the decisions of individuals to those of several other individuals. To see this, let us define the average decision similarity $\bar{S}_i$ of individual i as the average percentage agreement of this individual with all other $N - 1$ individuals, that is, for individual i, we calculate the percentage agreement with every other individual j ($j \neq i$) and take the average of these $N - 1$ percentages

$$E(\bar{S}_i) = \frac{\sum_{j=1..N, j\neq i} (p_i \cdot p_j + (1 - p_i) \cdot (1 - p_j))}{N - 1} \quad (3)$$

As above, we compare the expected decision similarity of two individuals j and i, which is given by

$$E(\bar{S}_j) - E(\bar{S}_i) = (p_j - p_i) \cdot \frac{N - 2}{N - 1} \cdot (2 \cdot \bar{p}_{-} - 1) \quad (4)$$

where $\bar{p}_{-}$ is the average decision accuracy of all individuals in the pool after excluding individuals j and i. As can be seen from Eq. 4, and analogous to our result above, whenever the remaining pool of individuals (i.e., excluding individuals j and i) is, on average, more often correct than incorrect (i.e., $\bar{p}_{-} > 0.5$), the individual with the higher accuracy is characterized by a higher expected average decision similarity, that is

$$p_j > p_i \Leftrightarrow E(\bar{S}_j) > E(\bar{S}_i) \quad (5)$$

Thus, for any pool of candidate decision makers where the average accuracy of individuals is above 0.5 (more precisely, where the average accuracy of individuals remains above 0.5 after excluding every possible pair of individuals), the higher an individual's average decision similarity, the more accurate this individual's decisions are. Figure 1B illustrates this predicted relationship for a pool of 10 individuals with decision accuracies of 0.55, 0.6, 0.65,…1.0, respectively.

Numerical simulations: Small samples and correlated decisions

The above result suggests that the similarity of an individual's decisions to the decisions of others—which can be readily observed when confronting individuals with a series of decision problems—is a powerful proxy for an individual's decision accuracy, provided the average accuracy of individuals exceeds 0.5. However, our analysis above is based on two simplifying assumptions. First, it is based on the expected value of decision similarity, that is, situations where for each individual a very large number of decisions are available for calculating that individual's decision similarity. Second, we assumed that decisions of different individuals are statistically independent from each other, that is, for any particular case, the accuracy p_i of individual i is independent of any other individual $j \neq i$ being correct or incorrect. To investigate the robustness of the similarity-accuracy relationship, we performed numerical simulations across a broad range of statistical environments relaxing these assumptions. We focused on different populations of decision makers (i.e., populations differing in their accuracy distribution) using the beta distribution, systematically varying the α and β shape parameters between 1 and 10 (Fig. 2A). These populations differ in their average accuracy, variance, and skewness, encompassing a wide variety of accuracy distributions. For each of these populations, we repeatedly and randomly sampled groups of 10 decision makers (i.e., individuals characterized by a given accuracy level) and confronted these individuals with a number of decision cases. We then investigated the strength and direction of the correlation between the observed decision similarity and decision accuracy,

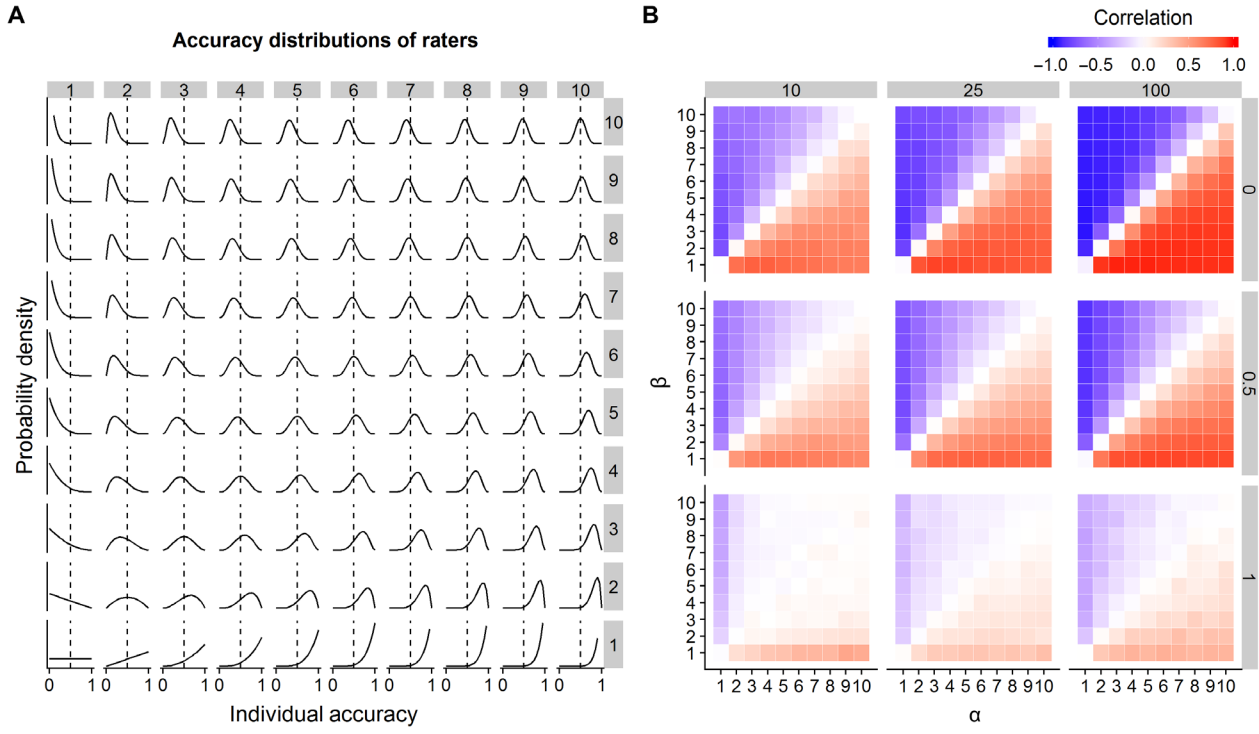

Fig. 2. Numerical simulations: The similarity-accuracy relationship is observed when similarity is calculated from a few samples and the decisions of different individuals are correlated with each other. (**A**) For the numerical simulations, we sampled decision makers from a wide range of populations of decision makers differing in their performance distribution (*x* axis, individual accuracy; *y* axis, probability density). We created those by systematically varying the two shape parameters α (values on top) and β (values on the right) of the beta distribution. Dashed vertical lines indicate the chance level of raters (i.e., accuracy of 0.5). (**B**) Average correlation coefficient between decision similarity and accuracy for 10 raters making 10, 25, and 100 decisions (subpanel columns) and for different degrees of correlations (0, 0.5, and 1; subpanel rows) between the decisions of different individuals (see Materials and Methods). Within each subpanel, the tiles correspond to raters drawn from the population (i.e., accuracy distribution) of the associated α-β combination in (A). Tiles below (above) the diagonal correspond to populations with an average individual accuracy above (below) 0.5; increasingly red (blue) colors indicate increasingly positive (negative) correlations. All results are averages over 2500 random samples. Whenever individual accuracy is above 0.5, we find a positive correlation between similarity and accuracy. While this correlation can be observed even in the most extreme scenarios with maximum correlation between the decisions of different decision makers (bottom row), generally, the strength of this correlation increases as the correlation between the decisions of decision makers decreases.

while varying (i) the number of observations used to calculate decision similarity (10, 25, and 100) and (ii) the degree to which the decisions of different individuals are correlated with each other (0, 0.5, and 1.0; see Materials and Methods).

Figure 2B shows the results of this analysis. Within each subpanel, the different tiles correspond to the different populations (i.e., accuracy distributions) of the associated tiles in Fig. 2A. While tiles below the diagonal (i.e., $\alpha > \beta$) correspond to populations with an average individual accuracy above 0.5, tiles above the diagonal (i.e., $\alpha < \beta$) correspond to populations with an average individual accuracy below 0.5. As can be seen from Fig. 2B, independent of the specific accuracy distribution, whenever the average individual accuracy in the population exceeds 0.5 (tiles below the diagonal), we find a positive correlation between decision similarity and accuracy. As expected, the strength of this correlation increases as (i) more observations are available to calculate decision similarity (moving from the left to the right panels), and (ii) decisions of individuals are less correlated with each other (moving from the bottom to the top panels). The correlation between decision similarity and accuracy can be observed in almost all scenarios (with the exception of the most extreme scenario of few samples and maximally correlated decisions). Our analysis, thus, strongly suggests that the similarity-accuracy relationship is robust.

Despite having considered a broad range of conditions in our numerical simulations, these simulations do not cover all conceivable scenarios. For simplicity, we have assumed that correlations do not differ systematically between subgroups of individuals and/or cases. In the presence of these systematic differences, scenarios are conceivable where the positive correlation between decision similarity and accuracy will not be observed, even when the average accuracy of individuals exceeds 0.5 (see the Supplementary Materials for an example).

Empirical analysis: Identifying high-performing individual decision makers

Our analysis above suggests that selecting decision makers based on their decisions' similarity to those of others should be a powerful and broadly applicable approach to identifying high performers. To test this prediction in real-world contexts, we analyzed data from several published datasets from three domains: medical diagnostics, geopolitical forecasting, and general knowledge. In particular, we investigated (i) a breast cancer dataset comprising 15,655 diagnoses by 101 radiologists based on 155 mammograms (*17*); (ii) a skin cancer dataset comprising 4320 diagnoses by 40 dermatologists based on 108 dermoscopic images of skin lesions (*18*); (iii) a geopolitical forecasting dataset from the Good Judgment Project containing 8460 forecasts

by 90 forecasters of 94 geopolitical events (*19*); and (iv) a dataset on general knowledge questions (here, which of two cities is larger) containing 99,000 decisions by 99 individuals on 1000 questions (*20*). For the medical datasets, the patient's actual health state (i.e., cancer present versus absent) was known from follow-up research (see Materials and Methods). Similarly, for the forecasting dataset, the correctness of the forecasts was determined from follow-up research. Figure S1 shows that in each of the datasets, (i) decision makers differ substantially in their individual performance, (ii) average individual accuracy is substantially above chance level, and (iii) decisions of different individuals are moderately correlated.

To test our basic prediction that individuals' decision similarity is positively correlated with their decision accuracy, we created, within each dataset, all possible pairs of decision makers and calculated, for each pair, their percentage agreement. As in our analysis above, each individual's average decision similarity was defined as its average percentage agreement with all other individuals. As predicted, we found a strong positive correlation in all four datasets between an individual's average decision similarity and accuracy (Fig. 3; Spearman's rank correlations: breast cancer: $r_s = 0.56$, $P < 0.001$; skin cancer: $r_s = 0.83$, $P < 0.001$; geopolitical events: $r_s = 0.84$, $P < 0.001$; city demographics: $r_s = 0.84$, $P < 0.001$). The network graphs in fig. S2 illustrate that while high performers are similar to other high performers, low performers are not similar to other low performers. That is, high performers perform well in the same way, whereas low performance is poor in myriad ways. Figure S3 zooms in on the case level, showing that—in line with our results above (Eq. 4)—high similarity is associated with high accuracy for cases in which the majority of the population is correct (a.k.a. "kind" cases), but not for cases in which the minority of the population is correct (a.k.a. "wicked" cases). Figure S4 shows that the positive relationship between similarity and accuracy is also observed when using the continuous probability scale in the forecasting dataset (rather than the binary yes/no scale).

To test whether the correlation between decision similarity and accuracy can be exploited to predict which individuals are high performers, we completed cross-validation procedures using training and test sets. In the training set, we calculated the similarity measure as above and selected individuals on that basis. We then looked at the performance of the selected individuals in the test set. To investigate how the number of decisions used to calculate decision similarity and the selection criterion (i.e., threshold of decision similarity) affected the selected individuals' performance, we repeated this analysis for training sets ranging from 0 to 60 cases in size and using selection criteria ranging from including individuals whose decision similarity was among the top 50%, top 25%, or top 5% in the training set (see Materials and Methods). Figure 4 shows the results of this analysis: In each of the three domains (medical diagnostics, geopolitical forecasting, and general knowledge), selecting individuals based on decision similarity is a powerful way to select high performers. Even when evaluating only a small number of cases in the training set and using a lenient selection criterion (e.g., top 50%), our approach succeeds in selecting individuals who perform substantially above average in the

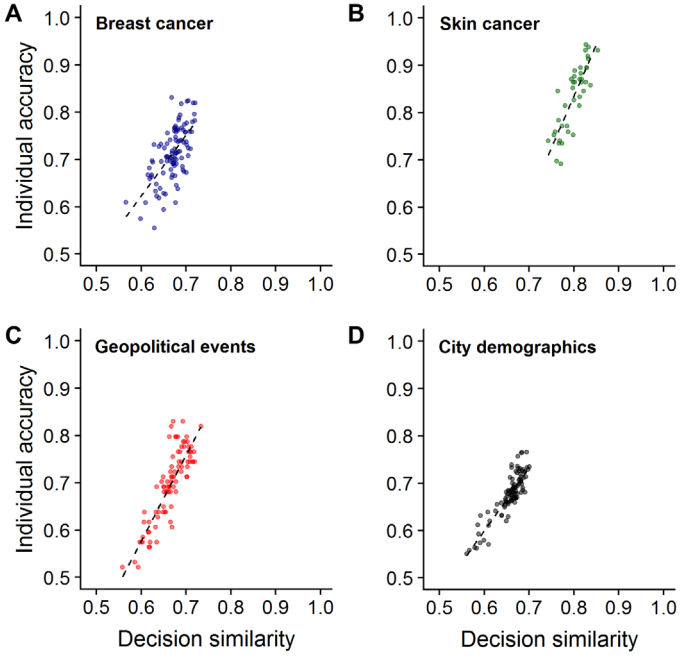

Fig. 3. Decision similarity tightly correlates with decision accuracy in breast and skin cancer diagnostics, geopolitical forecasting, and a general knowledge task. (**A** to **D**) In all four datasets, we find, as predicted (Figs. 1 and 2), a positive relationship between individuals' average decision similarity (i.e., average percentage of agreement with others) and accuracy. In (A) and (B), accuracy is expressed as balanced accuracy, and in (C) and (D), as proportion correct (see Materials and Methods). Lines are robust linear regression lines.

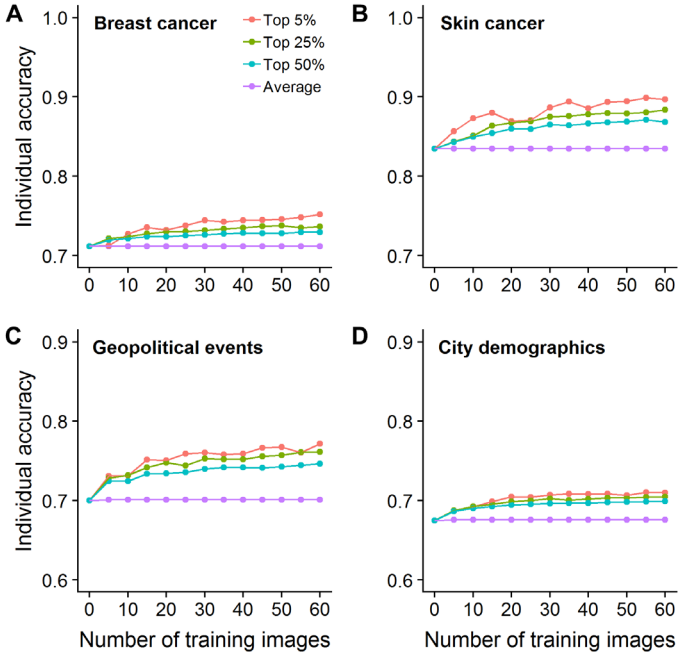

Fig. 4. Decision similarity robustly permits the identification of high-performing individuals. (**A** to **D**) The average performance of individuals in a test set selected on the basis of their decision similarity in a training set, for different decision similarity thresholds (e.g., the top 25% corresponds to the 25% of raters with the highest decision similarity in the training set) and different numbers of training images (i.e., number of decisions used to calculate decision similarity). As can be seen, in all four datasets, selecting individuals based on decision similarity substantially increases the average performance in the pool of decision makers. As predicted, when increasing the size of the training set and/or applying a stricter selection criterion, the average accuracy of the selected individuals increases. In (A) and (B), accuracy is expressed as balanced accuracy, and in (C) and (D), as proportion correct.

test set. Moreover, as we increase the size of the training set and/or apply a stricter selection criterion, the average accuracy of the selected individuals increases. Figure 4 focuses on selecting a subset of high performers from a relatively large pool of individuals (101 for breast cancer, 40 for skin cancer, 90 for geopolitical events, and 99 for city demographics). The same approach can be used to identify (relatively) low performers (fig. S5) and to select high and low performers from pools as small as three individuals (fig. S6).

Wisdom of crowds: Identifying high-performing groups

The similarity-accuracy relationship also has important implications for harnessing the "wisdom of crowds" when pooling multiple decision makers' independent decisions pertaining to the same case (*21*–*23*). One of the most commonly used aggregation rules is the majority rule (*24*, *25*), which is known to outperform individual decision-making in a variety of contexts, including medical diagnostics and forecasting (*7*, *26*–*30*). All other things being equal, the performance of the majority rule will increase with the average accuracy of the decision makers in that group—provided that individuals are, on average, more often correct than not (*25*, *31*). Consequently, our analysis above suggests that under a majority rule, groups consisting of individuals with relatively high decision similarity should outperform groups with relatively low similarity, because the individuals' decision similarity is positively correlated to their accuracy. Figure 5A illustrates this prediction, using a hypothetical example of eight groups of three identical decision makers with accuracies of 0.60, 0.65,…0.95, respectively.

To test this prediction pertaining to groups empirically, within each dataset, we randomly sampled groups of three individuals and combined their decisions using the majority rule (i.e., we selected, for each case, the decision that received either two or three votes). We then investigated how the average decision similarity among group members affected their collective accuracy. Consistent with our prediction, we find a positive relationship between decision similarity and collective accuracy (filled dots, Fig. 5, B to E). This effect is driven by the strong positive relationship between the decision similarity of group members and the average individual accuracy (open dots, Fig. 5, B to E). As shown in fig. S7, these results generalize to larger group sizes.

To test whether this relationship can be exploited to identify high-performing groups of individuals from a larger pool of groups, we again ran cross-validation procedures using training and test sets (see Materials and Methods). Figure 6 shows the results: In each of the three domains, selecting groups of individuals with a high-average decision similarity is a powerful method of identifying high-performing groups. Mirroring our results above, as we increase the size of the training set and/or apply a stricter selection criterion, the accuracy of the groups increases. While Fig. 6 focuses on identifying high-performing groups, the same approach can be used to identify low-performing groups (fig. S8).

Similarity to the majority

While our analyses above are based on the average decision similarity of individuals to others, a closely related approach is based on the similarity of individuals to the majority decision (i.e., the frequency of cases for which an individual makes the same decision as the majority). In the Supplementary Materials, we show that both approaches give rise to qualitatively the same analytical predictions, implying that

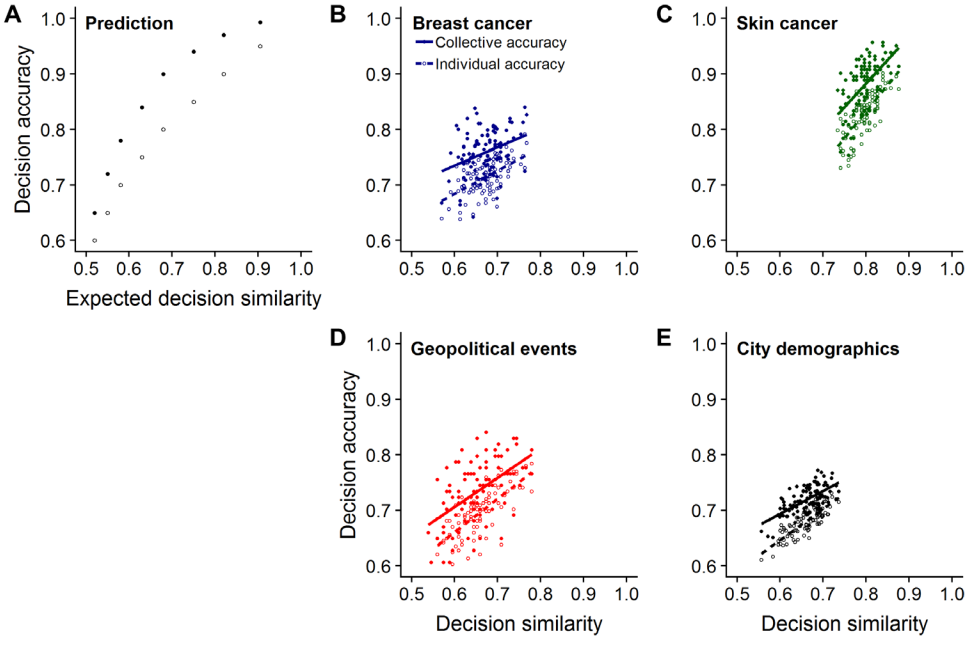

Fig. 5. The decision similarity within a group is tightly associated with that group's collective accuracy under the majority vote. (**A**) Illustrative example of the relationship between the expected average decision similarity, the average individual accuracy of group members (open dots), and the expected performance of the majority rule (filled dots), for eight groups of three identical decision makers with accuracies of 0.60, 0.65,…0.95, respectively. Decision similarity is calculated using Eq. 3, and the expected accuracy of the majority rule is calculated using the binomial distribution given by $p^3 + 3 \cdot p^2 \cdot (1 - p)$. (**B** to **E**) As predicted, in all datasets, we find a strong positive correlation between the average decision similarity among group members and their collective performance under the majority rule (filled dots and solid robust regression lines). This pattern is driven by a strong positive relationship between the average decision similarity among group members and the average individual performance of group members (open dots, dashed robust regression lines). In (B) and (C), accuracy is expressed as balanced accuracy, and in (D) and (E), as proportion correct.

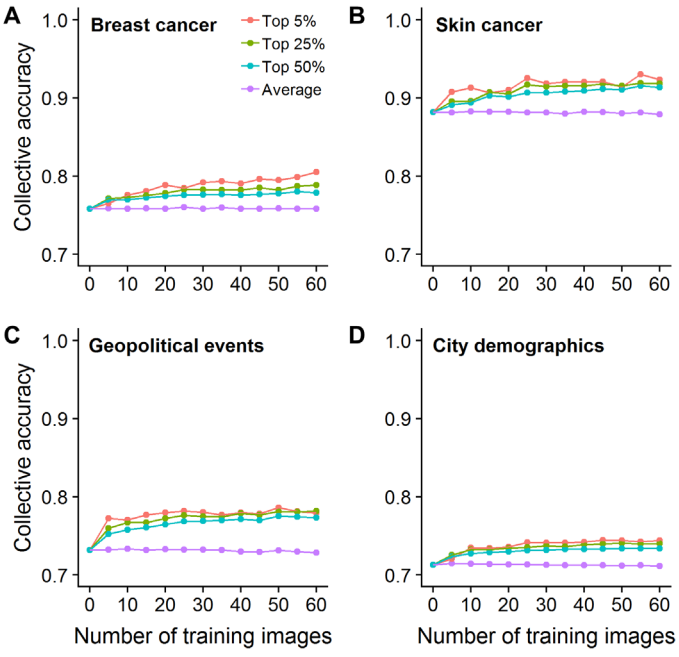

Fig. 6. Decision similarity permits identification of high-performing groups of individuals. (**A** to **D**) The average collective performance of groups in a test set, using the majority rule, when groups of individuals are selected on the basis of decision similarity in a training set, for different similarity thresholds (e.g., the top 25% corresponds to groups containing individuals with the 25% highest decision similarity values) and different numbers of training images (i.e., number of decisions used to calculate decision similarity). As can be seen, in all datasets, selecting groups of individuals based on decision similarity substantially increases the average collective performance. As predicted, the performance of the selected groups of individuals in the test set increases with the number of training images as well as with a stricter threshold value. The purple line ("Average") refers to the average majority rule performance of all groups in the test set. In (A) and (B), accuracy is expressed as balanced accuracy, and in (C) and (D), as proportion correct.

qualitatively, the same empirical results are to be expected when selecting individuals based on their similarity to the majority decision. In line with this, fig. S9 shows that, in each of the four datasets, there is a tight positive correlation between an individual's average decision similarity to others and that individual's similarity to the majority decision. Thus—as in the case of average decision similarity to others—there is a positive correlation between an individual's similarity to the majority decision and that individual's accuracy (fig. S10), and this relationship also permits the identification of high-performing individuals and groups.

Note that selecting individuals with the highest similarity values—be it the average decision similarity to others or the similarity to the majority—does not imply that these approaches aim to approximate the performance of the majority rule. To see this, note that both similarity measures tightly correlate with decision accuracy. Selecting the most similar individuals thus aims at selecting the most accurate individuals which—dependent on the distribution of accuracies and the group size—may perform systematically better or worse than the majority rule (see the Supplementary Materials for examples).

DISCUSSION

To summarize, on the basis of a general mathematical argument and numerical simulations, we predicted that in binary decision problems in which individuals are, on average, more often correct than not, (i) an individual's decision similarity to others is tightly correlated with that individual's decision accuracy, and (ii) this correlation can be used to identify high-performing individuals and groups of decision makers. Our analyses of four large datasets from three domains (medical diagnostics, geopolitical forecasting, and general knowledge) confirm these predictions. On the basis of these findings, we propose a widely applicable and robust method of identifying high-performing individuals and groups for binary decision problems. This method is captured by a simple and easily applicable heuristic: Faced with the task of identifying high performers from a pool of candidate decision makers (whose decision accuracy is unknown), observe a series of decisions of individuals and select those individuals (or groups) who make decisions that are most similar to others.

The practical use of our approach resides in that our method relies solely on input that is readily observable and does not require any knowledge about individuals' past performance. However, it does require that the average individual accuracy in the pool of candidate decision makers exceeds 50% (see Eq. 4 and Fig. 2). Arguably, it is a reasonable expectation that in many, if not most, binary decision problems, this condition is fulfilled. Moreover, while information on individuals' relative performance can be noisy, difficult, and costly to obtain—or unavailable—our approach requires only population-level information about whether or not individuals in the pool of candidate decision makers are, on average, more often correct than not. For practical implementation, we have two concrete recommendations. First, since our method aims at distinguishing between high- and low-performing individuals and groups, the benefits of our approach increase with larger expected performance differences between individuals in a domain. Second, in all four datasets, we found that observing around 20 to 30 decisions was enough to reliably identify high-performing individuals (Fig. 4) and groups (Fig. 6), as well as low-performing individuals and groups (figs. S5 and S8, respectively). That is, one may often not need more decisions to reliably categorize individuals and groups.

The positive relationship between similarity to others and accuracy can break down when correlations differ systematically between different subgroups of individuals and/or cases (see the Supplementary Materials for an example). The datasets we investigated, however, did not provide any evidence for this. In addition, in our datasets, the tight positive relationship between decision similarity to others and accuracy is only observed for cases in which the majority is correct (fig. S3). For these "kind" cases, individuals making the correct decision also had higher decision similarity. However, for "wicked" cases, we did not observe this relationship, showing that—while our approach does select high-performing individuals—it does not succeed in selecting individuals that solve these more difficult cases. Prelec *et al.* (*32*) showed that identifying answers that are more popular than people predict can help in solving these wicked cases.

In our mathematical and numerical analyses, we have assumed that each individual decision maker is characterized by a single number corresponding to the probability to be correct in any particular case. Although this is a common assumption made for analytical tractability, in many binary decision problems, the world is in one of two states (*33*, *34*) (e.g., cancer present versus cancer absent) and the accuracy of an individual may differ between these states (e.g., sensitivity versus specificity of a decision maker). A more general framework would, thus, characterize each individual by multiple

accuracies corresponding to the accuracies in each of these states. It will be interesting to investigate the consequences of such a more general conceptualization for the here developed ideas.

It is widely held that groups must harbor high diversity to maximize collective performance (*35*, *36*). When selecting members of decision-making groups, it is worth remembering that diversity is multidimensional and complex, encompassing individual differences on dimensions ranging from gender, ethnicity, age, and educational background to cognitive diversity and the correlation of errors among decision makers (*37*, *38*). Although some dimensions of diversity are known to boost collective accuracy (*35*, *36*), other dimensions can have negative consequences for collective performance, as illustrated here. Counterintuitively, selecting individuals who make similar decisions collectively outperforms selecting individuals whose decisions are dissimilar. Bahrami and colleagues (*39*, *40*) recently showed that two heads collectively outperform the ability of any single pair member whenever the pair members are of similar visual sensitivity [see also (*7*)]. Whether (and when) decision similarity reflects similarity in visual sensitivity or, perhaps, can complement it is a promising avenue for future research.

A related question is what underlies the observed differences in agreement level between decision makers. Einhorn (*41*) distinguishes between agreement "in fact" and agreement "in principle," the former referring to agreement of actual decisions, whereas the latter has to do with agreement about information and how it should be weighted in formulating a decision. Since we only have evaluated agreement "in fact," we cannot evaluate the extent to which the decision makers would agree in principle, and it will be interesting to investigate this issue in future research. Future work could also focus on the question to what extent the differences in agreement are the direct result of differences in individual accuracies (the approach we have followed here, but see the Supplementary Materials), or whether there are additional sources of covariation. For example, by using the κ statistic (instead of percentage agreement), which is a measure of interrater agreement that quantifies the agreement beyond that already expected by chance.

Our approach shares similarities to cognitive models based on cultural consensus theory (CCT) that use the agreement among individuals in answering a common set of questions to simultaneously estimate individuals' competences, response tendencies, and the "culturally correct" answers to the questions—by up-weighting the opinions of individuals with higher estimated competence (*42*–*45*). While a CCT approach is based on cognitive models aimed at describing the cognitive process involved in answering questions by estimating latent variables, our approach is substantially simpler: Instead of making assumptions about the cognitive process underlying decisions, it simply focuses on identifying individuals or groups with high future performance. Notwithstanding those differences, both approaches illustrate how agreement among people can be used to improve decision-making systems.

The finding that decision similarity can be harnessed to predict the future performance of individuals and groups is both powerful and simple. While we tested this prediction in medical diagnostics, geopolitical forecasting, and a general knowledge task, we expect similar results in other binary decision problems. Further developing and testing our approach in nonbinary decision-making contexts will be an important next step (see also fig. S4). We believe that the decision similarity-accuracy relationship offers a powerful approach to improve real-world systems of individual and collective decision-making.

MATERIALS AND METHODS

Numerical simulations

From each of the different populations of decision makers (Fig. 2A), we repeatedly sampled sets of 10 decision makers. Each of these 10 decision makers was characterized by a single value p_i indicating its average individual accuracy. Next, given this average accuracy, each of these decision makers evaluated M cases (10, 25, or 100). To illustrate, a decision maker with $p_i = 0.7$ evaluating 100 cases would be characterized by a vector of 100 values, where each value is either 0 (incorrect decision) or 1 (correct decision), drawn from a Bernoulli distribution with probability of 0.7 for a correct decision.

To study the effect of correlations between the decisions of different decision makers, we used an "opinion leader" approach (*25*). One of the 10 individuals was randomly assigned as the opinion leader, and we fixed the sequence of this individual's decisions (i.e., the sequence of 0s and 1s). Then, for all remaining individuals, the sequence of their decisions was paired to the opinion leader's sequence, depending on a correlation parameter p_c ($0 \leq p_c \leq 1$). In particular, starting at case $i = 1$, for each case, with probability $(1 - p_c)$, we randomly selected a decision from the set of remaining cases from that individual (i.e., decisions from cases j that have not yet been selected, $j \geq i$), and with probability p_c, we took the same decision as the decision of the opinion leader from this set. If the same decision was not present in the set of remaining decisions of that individual, we randomly selected a decision from this set. We then moved on to the next case $i + 1$. This procedure, thus, introduces different levels of correlation between decision makers, ranging from 0 (maximum amount of independence) to 1 (maximum amount of dependence) while not changing the frequency of 0s and 1s for each decision maker. Note that even if $p_c = 1$, there can still be disagreement between a pair of raters, namely, when the numbers of 0s and 1s in their respective vectors are not equal.

Next, we calculated, for each individual, his or her average percentage of agreement with the other nine decision makers over all M cases. Last, we calculated the Spearman's rank correlation coefficient between the average percentage agreement and average individual accuracy (p_i) across the 10 decision makers. For each unique combination of (i) number of cases M, (ii) level of correlations p_c, and (iii) population accuracy distribution (Fig. 2A), we repeated this procedure 2500 times, and we show values averaged across all repetitions (color codes in Fig. 2B).

Breast cancer dataset

The full information on the breast cancer dataset can be found in (*17*), and we summarized the dataset in (*7*). Therefore, we here provide a brief summary. Mammograms were randomly selected from screening examinations performed on women aged 40 to 69 between 2000 and 2003 from U.S. mammography registries affiliated with the Breast Cancer Surveillance Consortium (BCSC; Carolina Mammography Registry, New Hampshire Mammography Network, New Mexico Mammography Project, Vermont Breast Cancer Surveillance System, and Group Health Cooperative in western Washington). Radiologists who interpreted mammograms at facilities affiliated with these registries between January 2005 and December 2006 were invited to participate in this study, as were radiologists from Oregon, Washington, North Carolina, San Francisco, and New Mexico. Of the 409 radiologists invited, 101 completed all procedures and were included in the data analyses. Each screening examination included images from the current examination and one previous examination (allowing the radiologists to compare potential changes over time) and presented the craniocaudal and mediolateral oblique views of each breast (four views per woman

for each of the screening and comparison examinations). This approach is standard practice in the United States. Women who were diagnosed with cancer within 12 months of the mammograms were classified as patients with cancer ($n = 27$). Women who remained cancer free for a period of 2 years were classified as noncancerous patients ($n = 128$; 17% prevalence).

Radiologists viewed the digitized images on a computer (home computer, office computer, or laptop provided as part of the original study). All computers were required to meet all viewing requirements of clinical practice, including a large screen and high-resolution graphics ($\geq$1280 × 1024 pixels and a 1280-megabyte video card with 32-bit color). Radiologists saw two images at the same time (left and right breasts) and were able to alternate quickly ($\leq$1 s) between paired images, to magnify a selected part of an image, and to identify abnormalities by clicking on the screen. Each case presented craniocaudal and mediolateral oblique views of both breasts simultaneously, followed by each view in combination with its prior comparison image. Cases were shown in random order. Radiologists were instructed to diagnose them using the same approach they used in clinical practice (i.e., using the breast imaging reporting and data system lexicon to classify their diagnoses, including their decision that a woman be recalled for further examination). Radiologists evaluated the cases in two stages. In stage 1, four test sets were created, each containing 109 cases. Radiologists were randomly assigned to one of the four test sets. In stage 2, one test set containing 110 cases was created and presented to all radiologists. Some of the cases used in stage 2 had already been evaluated by some of the radiologists in stage 1. To avoid having the same radiologist evaluate a case twice, we excluded all cases from stage 2 that had already been viewed by that radiologist in stage 1. Moreover, we only included cases present in all four test sets to ensure that each radiologist evaluated the same set of cases, resulting in 155 unique cases. Between the two stages, radiologists were randomly assigned to one of three intervention treatments. Because there were no strong treatment differences (*46*), we pooled the data from stages 1 and 2. In our analysis, we treated the recommendation that a woman should be recalled for further examination as a positive test result.

The breast cancer data were assembled at the BCSC Statistical Coordinating Center in Seattle and analyzed at the Max Planck Institute for Human Development in Berlin, Germany. Each registry received institutional review board approval for active and passive consent processes or was granted a waiver of consent to enroll participants, pool data, and perform statistical analysis. All procedures were in accordance with the Health Insurance Portability and Accountability Act. All data were anonymized to protect the identities of women, radiologists, and facilities.

Skin cancer dataset

The full information on the skin cancer dataset can be found in (*18*), and we summarized the dataset in (*7*, *27*). Therefore, we here provide a brief summary. This dataset comprises 4320 diagnoses by 40 dermatologists of 108 skin lesions and were collected during a web-based consensus meeting (the Consensus Net Meeting on Dermoscopy). Skin lesions were obtained from the Department of Dermatology, University Federico II (Naples, Italy); the Department of Dermatology, University of L'Aquila (Italy); the Department of Dermatology, University of Graz (Austria); the Sydney Melanoma Unit, Royal Prince Alfred Hospital (Camperdown, Australia); and the Skin and Cancer Associates (Plantation, Florida). The lesions were selected on the basis of the photographic quality of the clinical and dermoscopic images available. The aim of the study was to diagnose whether or not a skin lesion was a melanoma, the most dangerous type of skin cancer. Histopathological specimens of all skin lesions were available and judged by a histopathology panel (melanoma, $n = 27$; no melanoma, $n = 81$; 25% prevalence). All participating dermatologists had at least 5 years of experience in dermoscopy practice, teaching, and research. They first underwent a training procedure in which they familiarized themselves with the study's definitions and procedures in web-based tutorials with 20 sample skin lesions. They subsequently evaluated 108 skin lesions in a two-step online procedure. First, they used an algorithm to differentiate melanocytic from nonmelanocytic lesions. Whenever a lesion was evaluated as melanocytic, dermatologists were asked to classify it as either a melanoma or a benign melanocytic lesion using four different algorithms. Here, we focus on the diagnostic algorithm with the highest diagnostic accuracy, which is also the one most widely used for melanoma detection: pattern analysis. It uses a set of global (textured patterns covering most of the lesion) and local features (representing characteristics that appear in part of the lesion) to differentiate between melanomas and benign melanocytic lesions. We treated the decision to classify a lesion as melanoma as a positive test result.

The review board of the Second University of Naples waived approval because the study did not affect routine procedures. All participating dermatologists signed a consent form before participating in the study.

Forecasting dataset

The forecasting dataset is part of the Good Judgment Project (*19*). This is a large-scale forecasting project running over several years and using a wide variety of participants and settings (e.g., training schedules and team competitions). We used data from the first year of the forecasting project (*47*). In this year, 102 questions, such as "Will Serbia be officially granted EU candidacy by 31 December 2011?" and "Will the Nikkei 225 index finish trading at or above 9500 on 30 September 2011?" had to be forecasted. Participants were asked to estimate the probability of the future event. We excluded questions with more than two possible answers and questions for which the correct answer could not be irrefutably determined ($n = 8$). We excluded forecasters who did not complete all remaining 94 questions, resulting in 90 forecasters. The total dataset we used thus contained 8460 forecasts by 90 forecasters on 94 geopolitical events. Sometimes, forecasters updated their forecasts over time, thereby giving multiple responses. In these cases, we used their first forecast only. To investigate the case of binary decision-making, we converted the probability scores into 0 (probabilities <0.5) and 1 (probabilities >0.5). Scores that were 0.5 were randomly converted to either 0 or 1. In fig. S4, we investigate a scenario when using the probability scores directly without any conversion. Participants were free to enter the forecasting competition, and the subject pool consisted of a mix of laypeople and geopolitical experts.

General knowledge dataset

This dataset is based on study 3 in (*20*) and contains binary responses to the question, "Which of the following two cities has more inhabitants?" The stimulus set consisted of 1000 randomly generated pairs of cities from a list of the 100 most populous cities in the United States in 2010 as determined by the U.S. Census Bureau. After observing a fixation cross, participants saw a pair of cities, and after 1.6 s, they were cued to make a decision. Participants rated 1000 pairs, distributed over two sessions. Participants ($n = 109$) were recruited from the Michigan State University (MSU) psychology research participant pool and received

class credits plus a $0 to $4 bonus per session. Informed consent was obtained from participants according to the guidelines of the MSU Institutional Review Board. We excluded participants who did not complete both sessions, resulting in 99 participants, all of whom provided a decision on each of the 1000 city pairs.

Empirical analysis: Individual decision makers

To study the relationship between individuals' decision similarity and decision accuracy, we calculated, within each dataset each individual's accuracy. For the forecasting and the general knowledge dataset, we used proportion correct. For the medical datasets, we used balanced accuracy, defined as (sensitivity + specificity)/2. This implies weighting the misclassification errors of positive and negative cases inversely proportional to their prevalence; in case of a 50% prevalence, balanced accuracy reduces to proportion correct. Using balanced accuracy circumvents the problem that proportion correct faces, namely, that when the prevalence becomes highly asymmetric, one can always perform well by always predicting the more frequent outcomes (e.g., if the prevalence of cancer is 10%, always predicting no cancer gives a proportion correct of 90%, but a balanced accuracy of 50%). Next, we calculated each individual's average decision similarity. For this, we used the percentage agreement, which ranges from 0 (never made the same decision for a case) to 1 (always made the same decision for all cases). To quantify the relationship between average decision similarity and individual accuracy (Fig. 3), we used Spearman's rank correlations (cor.test) in R (version 3.4.4).

To test whether decision similarity can be used to predict high-performing individuals, we performed a cross-validation procedure using a training and test set procedure. Within each dataset, we randomly drew m cases to create a training set (varying m from 0 to 60 in steps of 5). Part of the remaining cases were used for forming a test set. As size of the test set, we used the total number of cases in a dataset minus the maximum size of the training set (i.e., 60). This assured that, within each dataset, the test set size was the same across all number of training images, making the results comparable across training set sizes. In the training set, we calculated each individual's average decision similarity with all others (i.e., percentage agreement). We then ranked all individuals based on their decision similarity in the training set and tested their performance in the test set using different similarity thresholds (Fig. 4 and fig. S5). For example, in Fig. 4, the top 25% corresponds to individuals with the 25% highest decision similarity values in the training set. Similarly, in fig. S5, the bottom 25% corresponds to individuals with the 25% lowest decision similarity values in the training set. We repeated each number of training images 1000 times in each dataset and report the average values.

Empirical analysis: Groups of decision makers

To study the relationship between decision similarity of group members and collective accuracy, we randomly sampled, from each dataset, 100 unique groups of n individuals (n = 3 and 9). For each group, we calculated the (i) average individual decision similarity (i.e., the average percentage agreement of all possible pairwise combinations of group members), (ii) average individual accuracy of the group members, and (iii) accuracy of the majority rule (Fig. 5, B to E, and fig. S7).

To test whether decision similarity can be used to predict high-performing groups, we performed a cross-validation procedure using a training and test set procedure. Within each dataset, we randomly drew m cases to create a training set (varying m from 0 to 60 in steps of 5) and used part of the remaining cases to form a test set using the procedure described above. Within a training set, we calculated the average decision similarity of each rater and ranked individuals according to their decision similarity. We then created groups of three raters, calculated their average decision similarity, and tested the performance of these groups in the test set using different similarity thresholds (Fig. 6 and fig. S8). For example, in Fig. 6, the top 25% corresponds to groups with the 25% highest decision similarity values in the training set. We repeated each number of training images 1000 times in each dataset and report the average values.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/11/eaaw9011/DC1

Supplementary text

Fig. S1. Distribution of individuals' level of accuracy and correlated decisions in the four datasets.

Fig. S2. High-performing individuals are similar to each other, while low-performing individuals tend to make dissimilar decisions.

Fig. S3. Decision similarity performs well for cases in which the majority decided correctly but breaks down for cases in which the minority decided correctly.

Fig. S4. The similarity-accuracy relationship is also present when using the continuous probability forecasts.

Fig. S5. Decision similarity permits identification of low-performing individuals.

Fig. S6. Decision similarity permits identification of high-performing (and low-performing) individuals in small groups.

Fig. S7. The relationship between decision similarity of a group of nine individuals and their individual and collective accuracy.

Fig. S8. Decision similarity permits identification of low-performing groups.

Fig. S9. In each of the four datasets, the average decision similarity to others tightly correlates with the decision similarity to the majority judgment.

Fig. S10. Decision similarity to the majority tightly correlates with decision accuracy in breast and skin cancer diagnostics, geopolitical forecasting, and a general knowledge task.

Skin cancer data set

R Code numerical simulations (Fig. 2B)

REFERENCES AND NOTES

1. B. Mellers, E. Stone, T. Murray, A. Minster, N. Rohrbaugh, M. Bishop, E. Chen, J. Baker, Y. Hou, M. Horowitz, Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspect. Psychol. Sci.* **10**, 267–281 (2015).
2. M. Spann, H. Ernst, B. Skiera, J. H. Soll, Identification of lead users for consumer products via virtual stock markets. *J. Prod. Innov. Manage.* **26**, 322–335 (2009).
3. M. O'Sullivan, Unicorns or Tiger Woods: Are lie detection experts myths or rarities? A response to on lie detection "wizards" by Bond and Uysal. *Law Hum. Behav.* **31**, 117–123 (2007).
4. M. A. Burgman, M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, C. Twardy, Expert status and performance. *PLOS ONE* **6**, e22998 (2011).
5. B. Mellers, L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, P. Atanasov, S. A. Swift, T. Murray, E. Stone, P. E. Tetlock, Psychological strategies for winning a geopolitical forecasting tournament. *Psychol. Sci.* **25**, 1106–1115 (2014).
6. A. Rae, R. Alexander, Forecasts or fortune-telling: When are expert judgements of safety risk valid? *Saf. Sci.* **99**, 156–165 (2017).
7. R. H. J. M. Kurvers, S. M. Herzog, R. Hertwig, J. Krause, P. A. Carney, A. Bogart, G. Argenziano, I. Zalaudek, M. Wolf, Boosting medical diagnostics by pooling independent judgments. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8777–8782 (2016).
8. P. E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?-New Edition* (Princeton Univ. Press, 2017).
9. M. A. Burgman, *Trusting Judgements: How To Get The Best Out Of Experts* (Cambridge Univ. Press, 2016).
10. R. Cooke, *Experts In Uncertainty: Opinion And Subjective Probability In Science* (Oxford Univ. Press on Demand, 1991).
11. J. Witkowski, P. Atanasov, L. H. Ungar, A. Krause, Proper proxy scoring rules, *Thirty-First AAAI Conference on Artificial Intelligence* (2017).
12. N. Miller, P. Resnick, R. Zeckhauser, Eliciting informative feedback: The peer-prediction method. *Manage. Sci.* **51**, 1359–1373 (2005).
13. D. Prelec, A Bayesian truth serum for subjective data. *Science* **306**, 462–466 (2004).
14. H. Collins, R. Evans, *Rethinking Expertise* (University of Chicago Press, 2008).

15. A. Hart, *Knowledge acquisition for expert systems* (School of Computing, Lancashire Polytechnic, Preston, 1986).
16. J. Shanteau, D. J. Weiss, R. P. Thomas, J. C. Pounds, Performance-based assessment of expertise: How to decide if someone is an expert or not. *Eur. J. Oper. Res.* **136**, 253–263 (2002).
17. P. A. Carney, T. A. Bogart, B. M. Geller, S. Haneuse, K. Kerlikowske, D. S. Buist, R. Smith, R. Rosenberg, B. C. Yankaskas, T. Onega, D. L. Miglioretti, Association between time spent interpreting, level of confidence, and accuracy of screening mammography. *Am. J. Roentgenol.* **198**, 970–978 (2012).
18. G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, R. Hofmann-Wellenhof, M. Landthater, S. W. Menzies, H. Pehamberger, D. Piccolo, H. S. Rabinovitz, R. Schiffner, S. Staibano, W. Stolz, I. Bartenjev, A. Blum, R. Braun, H. Cabo, P. Carli, V. De Giorgi, M. G. Fleming, J. M. Grichnik, C. M. Grin, A. C. Halpern, R. Johr, B. Katz, R. O. Kenet, H. Kittler, J. Kreusch, J. Malvehy, G. Mazzocchetti, M. Oliviero, F. Özdemir, K. Peris, R. Perotti, A. Perusquia, M. A. Pizzichetta, S. Puig, B. Rao, P. Rubegni, T. Saida, M. Scalvenzi, S. Seidenari, I. Stanganelli, M. Tanaka, K. Westerhoff, I. H. Wolf, O. Braun-Falco, H. Kerl, T. Nishikawa, K. Wolff, Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *J. Am. Acad. Dermatol.* **48**, 679–693 (2003).
19. L. Ungar, B. Mellers, V. Satopää, P. Tetlock, J. Baron, The good judgment project: A large scale test of different methods of combining expert predictions, *AAAI Fall Symposium Series* (2012).
20. S. Yu, T. J. Pleskac, M. D. Zeigenfuse, Dynamics of postdecisional processing of confidence. *J. Exp. Psychol. Gen.* **144**, 489–510 (2015).
21. J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Knopf Doubleday Publishing Group, 2004).
22. A. E. Mannes, J. B. Soll, R. P. Larrick, The wisdom of select crowds. *J. Pers. Soc. Psychol.* **107**, 276–299 (2014).
23. S. M. Herzog, A. Litvinova, K. S. Yahosseini, A. Novaes Tump, R. H. J. M. Kurvers, The ecological rationality of the wisdom of crowds, in *Taming Uncertainty* (MIT Press, Cambridge, Massachusetts, 2019), pp. 245–262.
24. R. Hastie, T. Kameda, The robust beauty of majority rules in group decisions. *Psychol. Rev.* **112**, 494–508 (2005).
25. B. Grofman, G. Owen, S. L. Feld, Thirteen theorems in search of the truth. *Theor. Decis.* **15**, 261–278 (1983).
26. M. Wolf, J. Krause, P. A. Carney, A. Bogart, R. H. J. M. Kurvers, Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PLOS ONE* **10**, e0134269 (2015).
27. R. H. J. M. Kurvers, J. Krause, G. Argenziano, I. Zalaudek, M. Wolf, Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatol.* **151**, 1346–1353 (2015).
28. A. E. Murr, "Wisdom of crowds"? A decentralised election forecasting model that uses citizens' local expectations. *Elect. Stud.* **30**, 771–783 (2011).
29. R. H. J. M. Kurvers, A. de Zoete, S. L. Bachman, P. R. Algra, R. Ostelo, Combining independent decisions increases diagnostic accuracy of reading lumbosacral radiographs and magnetic resonance imaging. *PLOS ONE* **13**, e0194128 (2018).
30. A. Novaes Tump, M. Wolf, J. Krause, R. H. J. M. Kurvers, Individuals fail to reap the collective benefits of diversity because of over-reliance on personal information. *J. R. Soc. Interface* **15**, 20180155 (2018).
31. T. S. Wallsten, D. V. Budescu, I. Erev, A. Diederich, Evaluating and combining subjective probability estimates. *J. Behav. Decis. Mak.* **10**, 243–268 (1997).
32. D. Prelec, H. S. Seung, J. McCoy, A solution to the single-question crowd wisdom problem. *Nature* **541**, 532–535 (2017).
33. J. A. Marshall, R. H. Kurvers, J. Krause, M. Wolf, Quorums enable optimal pooling of independent judgements in biological systems. *eLife* **8**, e40368 (2019).
34. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
35. L. Hong, S. E. Page, Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16385–16389 (2004).
36. S. Krause, R. James, J. J. Faria, G. D. Ruxton, J. Krause, Swarm intelligence in humans: Diversity can trump ability. *Anim. Behav.* **81**, 941–948 (2011).
37. L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **51**, 181–207 (2003).
38. D. Bang, C. D. Frith, Making better decisions in groups. *R. Soc. Open Sci.* **4**, 170193 (2017).
39. B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, C. D. Frith, Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
40. D. Bang, R. Fusaroli, K. Tylén, K. Olsen, P. E. Latham, J. Y. F. Lau, A. Roepstorff, G. Rees, C. D. Frith, B. Bahrami, Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Conscious. Cogn.* **26**, 13–23 (2014).
41. H. J. Einhorn, Expert judgment: Some necessary conditions and an example. *J. Appl. Psychol.* **59**, 562–571 (1974).
42. A. K. Romney, S. C. Weller, W. H. Batchelder, Culture as consensus: A theory of culture and informant accuracy. *Am. Anthropol.* **88**, 313–338 (1986).
43. S. C. Weller, Cultural consensus theory: Applications and frequently asked questions. *Field Methods* **19**, 339–368 (2007).
44. S. C. Weller, N. C. Mann, Assessing rater performance without a "gold standard" using consensus theory. *Med. Decis. Making* **17**, 71–79 (1997).
45. M. Steyvers, B. Miller, Cognition and Collective Intelligence, in *Handbook of Collective Intelligence*, T. Malone, M. S. Berstein, Eds. (Cambridge, MA: MIT Press, 2015), pp. 119–125.
46. B. M. Geller, A. Bogart, P. A. Carney, E. A. Sickles, R. Smith, B. Monsees, L. W. Bassett, D. M. Buist, K. Kerlikowske, T. Onega, Educational interventions to improve screening mammography interpretation: A randomized controlled trial. *Am. J. Roentgenol.* **202**, W586–W596 (2014).
47. Good Judgment Project, GJP Data (Harvard Dataverse, V1, 2016).

Acknowledgments: We thank D. Ain, V. Chase, and two anonymous referees for detailed comments on an earlier version of this manuscript. We thank the Breast Cancer Surveillance Consortium (BCSC) for sharing the breast cancer dataset. We thank all participating women, facilities, and radiologists for the data they provided. For full acknowledgements, see https://bcsc-research.org/. **Funding:** Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 39052313. Collection of the breast cancer data was supported by the American Cancer Society using a donation from the Longaberger Company's Horizon of Hope Campaign (grants SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274-01, SIRSG-07-275, SIRSG-06-281, SIRSG-09-270-01, SIRSG-09-271-1, and SIRSG-06-290-04), by the Breast Cancer Stamp Fund, and by the National Cancer Institute Breast Cancer Surveillance Consortium (Contract HHSN261201100031C). **Author contributions:** R.H.J.M.K. and M.W. conceived and planned the paper with contributions from S.M.H., R.H., J.K., and M.M. G.A., I.Z., and P.A.C. performed the data collection. R.H.J.M.K. and M.W. performed the analysis. R.H.J.M.K. and M.W. wrote the paper with substantial input from all authors. **Competing interests:** The authors declare that they have no competing interest. **Data and materials availability:** The skin cancer dataset is included as a Supplementary Material. The geopolitical dataset is part of the Good Judgment Project and available online at dataverse.harvard.edu/dataverse/gjp. The general knowledge dataset is available online at https://osf.io/cuzqm/. The BCSC holds legal ownership of the breast cancer dataset. Information regarding data requests can be directed to the BCSC https://bcsc-research.org/. The code of the numerical simulations (Fig. 2) is included as a Supplementary Material. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 4 February 2019
Accepted 20 September 2019
Published 20 November 2019
10.1126/sciadv.aaw9011

Citation: R. H. J. M. Kurvers, S. M. Herzog, R. Hertwig, J. Krause, M. Moussaid, G. Argenziano, I. Zalaudek, P. A. Carney, M. Wolf, How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Sci. Adv.* **5**, eaaw9011 (2019).