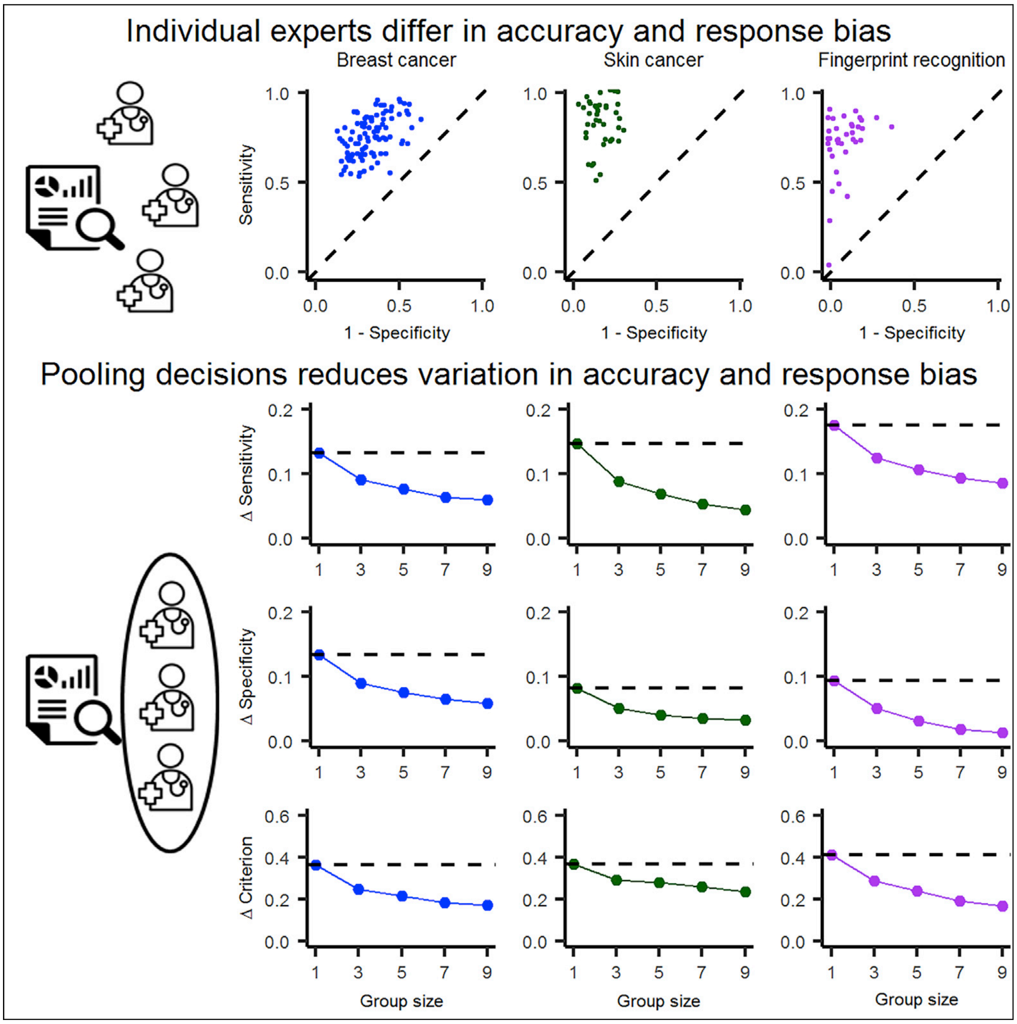
iScience

CellPress
OPEN ACCESS

Article

Pooling decisions decreases variation in response bias and accuracy



Ralf H.J.M. Kurvers, Stefan M. Herzog, Ralph Hertwig, Jens Krause, Max Wolf

kurvers@mpib-berlin.mpg.de

Highlights

Professional decision makers typically differ in their response bias and accuracy

Such differences undermine the reliability and fairness of decision systems

Pooling decisions reduces such variation in response bias and accuracy

This occurred in cancer diagnostics, fingerprint analysis, and forecasting

Kurvers et al., iScience *24*, 102740
July 23, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102740

iScience

CellPress
OPEN ACCESS

Article

Pooling decisions decreases variation in response bias and accuracy

Ralf H.J.M. Kurvers,[1,2,4,*] Stefan M. Herzog,[1] Ralph Hertwig,[1] Jens Krause,[2,3] and Max Wolf[2]

SUMMARY

Decision makers in contexts as diverse as medical, judicial, and political decision making are known to differ substantially in response bias and accuracy, and these differences are a major factor undermining the reliability and fairness of the respective decision systems. Using theoretical modeling and empirical testing across five domains, we show that collective systems based on pooling decisions robustly overcome this important but as of now unresolved problem of experts' heterogeneity. In breast and skin cancer diagnostics and fingerprint analysis, we find that pooling the decisions of five experts reduces the variation in sensitivity among decision makers by 52%, 54%, and 41%, respectively. Similar reductions are achieved for specificity and response bias, and in other domains. Thus, although outcomes in individual decision systems are highly variable and at the mercy of individual decision makers, collective systems based on pooling decrease this variation, thereby promoting reliability, fairness, and possibly even trust.

INTRODUCTION

Designing the process of making decisions well is key to the success of human societies, including good decision making in medicine, law, economics, and politics. The architecture of this process and the resulting decision-making systems are thus of paramount importance to human welfare. One of the prominent properties of decision-making systems is whether the system rests on an individual agent making the decision or on a collective of individuals. In recent decades, a large body of research across the behavioral sciences has sought to map out the relative advantages and disadvantages of individual versus collective decision-making systems (Bang and Frith, 2017; Conradt and List, 2009; El Zein et al., 2019; Gully et al., 2002; Kerr and Tindale, 2004). To date, this research's focus has been almost exclusively on accuracy as the performance criterion (Bahrami et al., 2010; Bang and Frith, 2017; Clément et al., 2013; Kerr and Tindale, 2004; Koriat, 2012; Kurvers et al., 2016, 2019; Lorenz et al., 2011; Wolf et al., 2013; Woolley et al., 2010) (but see El Zein et al., 2019), neglecting a second major performance metric of the quality of decision-making systems, namely, the variation in outcomes between decision-making agents. Here, we turn to this quality benchmark. We start with explaining the challenge heterogeneity in expert performance poses for good decisions.

Individual decision makers differ in key aspects of their decision processes and thus the decisions they produce. In many domains including medical diagnostics, truth detection, and geopolitical forecasting, experts have been shown to differ substantially in their level of accuracy (Burgman, 2016; Koran, 1975; Kurvers et al., 2016; Mellers et al., 2015; O'Sullivan and Ekman, 2004; Wolf et al., 2015). Moreover, in binary classification tasks, such as truth detection and many instances of medical diagnostics, experts have also been shown to differ in how much evidence they require to classify a case as either signal (e.g., cancer, lie) or noise (e.g., non-cancer, truth). In other words, they differ in their response bias (DeKay, 1996; Deneef and Kent, 1993; Hammond, 1996; Macmillan and Creelman, 2005). These differences matter because, in general, the higher the variation in outcomes (i.e., accuracy levels, response bias) between decision makers, the less predictable and reliable the outputs of the corresponding decision-making system. Systems with low variation are perceived as fairer and more trustworthy than those in which outcomes depend heavily on the specific decision makers (Tyler, 2000). To appreciate the issue of fairness, consider the principle of equality before the law. This fundamental legal principle is typically taken to imply that judicial outcomes should not depend on irrelevant characteristics (e.g., race or gender) of the defendant, yet it also implies that outcomes should be independent of the legal agent (e.g., court or jury) deciding on a defendant's case. Similarly, in the medical domain, patients facing a health crisis hope that the diagnosis and

[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] Faculty of Life Sciences, Albrecht Daniel Thaer-Institute of Agricultural and Horticultural Sciences, Humboldt-Universität zu Berlin, Invalidenstrasse 42, 10115 Berlin, Germany

[4] Lead contact

*Correspondence: kurvers@mpib-berlin.mpg.de

https://doi.org/10.1016/j.isci.2021.102740

treatment they receive is not subject to the idiosyncrasies of the physician they happen to encounter. Thus, in many contexts, next to high decision accuracy, a low degree of variation in outcomes is an important goal for decision-making systems.

In the following, we use a combined theoretical and real-world data-driven approach to investigate how a simple, yet highly effective, system of collective decision making—the pooling of independent decisions using the majority rule—affects the variation in outcomes between decision makers. Our focus is on binary classification tasks such as categorizing a mammogram in terms of ''cancer present'' or ''cancer absent.'' A first intuition can be derived from the law of large numbers: as group size becomes sufficiently large, the mean characteristics of the members of a group (e.g., the mean accuracy level or mean response bias of a random sample of decision makers) tend to approach the population mean, thereby reducing variation between different large groups. Here, however, we are interested in variation in decision outcomes between groups with sizes that are realistic in real-world decision-making environments in which groups and teams are often relatively small. In addition, and limiting the usefulness of the law of large numbers in this context, decision outcomes that stem from pooling independent decisions can generally not be predicted from the mean characteristics of the group's members alone.

We proceed in two steps. First, based on a statistical argument and large-scale numerical computer simulations, we derive general predictions about how the pooling of decisions affects variation in decision outcomes between decision-making agents in binary classification tasks. Second, we use several real-world datasets on breast and skin cancer diagnostics, fingerprint analysis, geopolitical forecasting, and a general knowledge task to empirically test these predictions.

RESULTS

Analytical result: Pooling decisions reduces outcome variation in homogeneous groups

We start by developing a basic formal intuition about the relationship between pooling decisions and outcome variation in binary classification tasks. The task is to solve a set of binary choice problems, for example, to classify a mammogram as "cancer present" or "cancer absent," to decide whether a defendant is guilty or innocent, or to predict whether or not an earthquake will occur in a given region. The most basic scenario is one in which individual decision makers do not differ in accuracy levels or response bias. We assume a population of individuals, with each one being characterized by the same accuracy level *a*, corresponding to their probability of being correct in any given choice. We further assume that decisions of different individuals are statistically independent from each other. Although individuals have identical accuracies, they may still differ in how they respond to specific cases. The expected frequency of disagreement between two individuals can be calculated from the binomial distribution as $2 \cdot a \cdot (1 - a)$. Focusing on situations where individuals' accuracy level is above chance (i.e., $a > 0.5$), we make two observations. First, the higher the accuracy *a* of two individuals, the lower their expected frequency of disagreement (the derivative of the expected disagreement with respect to *a* is $2 - 4 \cdot a$, which is negative as long as $a > 0.5$). Second, groups that pool independent decisions with the help of a majority rule achieve higher accuracy than individuals, with larger groups achieving higher accuracy than smaller groups (Condorcet Jury Theorem) (Grofman et al., 1983). From these two observations follows that—in this most basic scenario—pooling decisions decreases outcome variation between agents: as group size increases, accuracy increases, thereby decreasing the expected frequency of disagreement between groups. Figure 1 illustrates this relationship for different individual accuracy levels and group sizes. As can be seen, for any individual accuracy level *a*—with the exception of accuracies 0.0, 0.5, and 1.0—the variation between agents (1) is larger between individual decision makers than between groups of decision makers; (2) decreases as group size increases; and (3) approaches 0 as group size becomes sufficiently large. This general statistical argument concerning the most basic case (i.e., absence of individual differences in accuracy and response bias) suggests that pooling decisions could be a powerful approach to reduce variation in outcomes between decision-making agents. Note that individuals' accuracy level needs to be above chance ($a > 0.5$), so that pooling decisions simultaneously decreases outcome variation and increases accuracy. If accuracy is below chance ($a < 0.5$), pooling decisions will decrease both outcome variation and accuracy (see discussion).

Numerical simulations: Pooling decisions reduces variation in accuracy in heterogeneous groups

We next investigate how pooling decisions impacts variation in outcomes in more realistic scenarios, namely, when decision makers differ in accuracy and response bias. Note that, whenever individuals differ in accuracy and/or response bias, then the decision outcomes produced by a group employing the majority

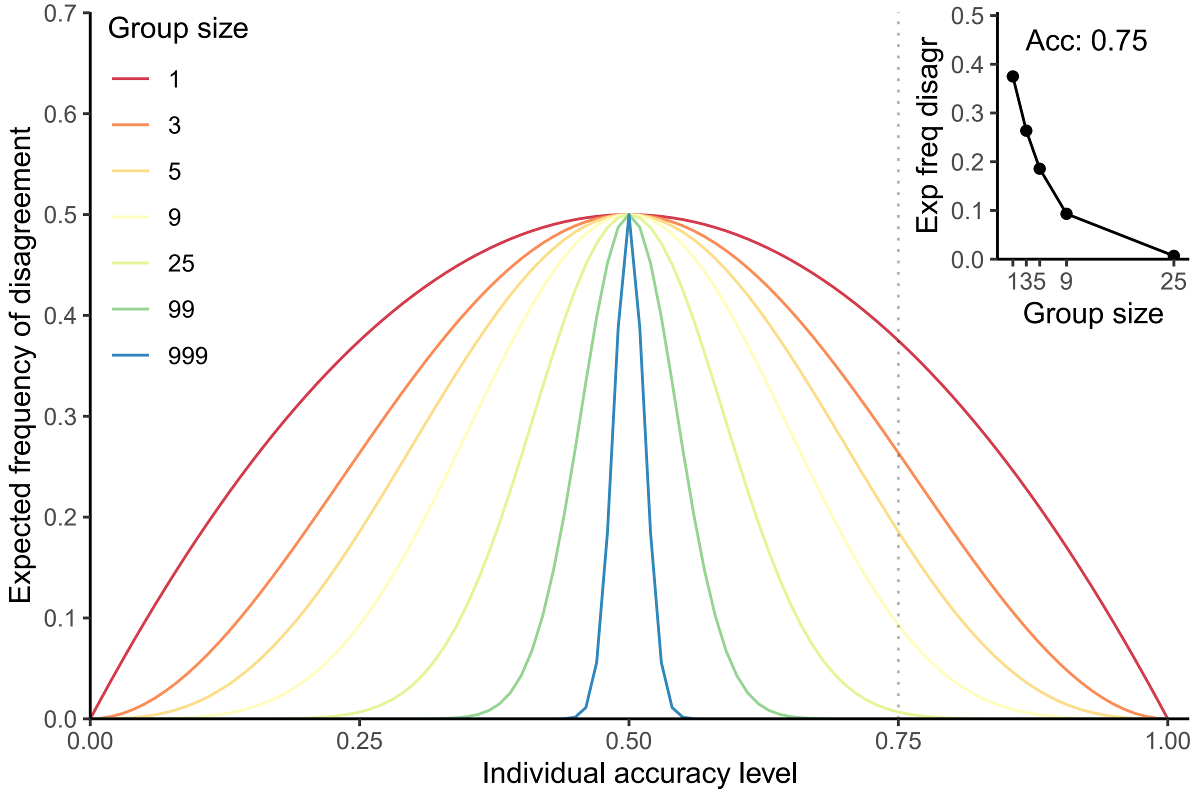


Figure 1. Statistical argument: Pooling independent decisions reduces outcome variation in homogeneous groups

We here consider the most basic scenario, where individual decision makers do not differ in accuracy or response bias. Although individuals have identical accuracies, they may differ in how they respond to specific cases. The expected frequency of disagreement between two decision-making agents (i.e., two individuals, two groups pooling independent decisions with a majority rule) can be calculated from the binomial distribution (see results). For any level of individual accuracy (x axis), the figure shows the expected frequency of cases where two agents disagree. With the exception of accuracies 0.0, 0.5, and 1.0, pooling decisions is predicted to systematically reduce the frequency of cases with disagreement between agents, with larger groups showing larger reductions than smaller groups. This reduction can be explained by (1) the link between accuracy levels and the expected frequency of disagreement and (2) the effect of pooling decisions on accuracy levels: as long as the individual accuracy is above chance (i.e., > 0.5), pooling decisions increases accuracy levels, which, in turn, decreases the expected frequency of disagreement. When the individual accuracy level drops below chance (i.e., < 0.5), pooling decisions decreases both the expected frequency of disagreement and accuracy. Only odd group sizes are calculated to avoid the need for a tie-breaking rule. The inset shows how expected disagreement develops with group size for an individual accuracy level of 0.75.

rule (and thus the variation in outcomes between groups) will depend on the specific composition of individuals in that group. To investigate this more complex scenario we will employ numerical simulations.

We first focus on differences in accuracy levels. To this end, we consider a large range of populations of decision makers that differ in their accuracy distribution. To generate these distributions, we use beta distributions—the canonical method to model proportions—which are defined on the interval [0, 1], thus matching the natural accuracy range of 0 (always incorrect) to 1 (always correct). We systematically vary the distributions' mean and variance (controlled by the two shape parameters α and β), resulting in populations that differ in their average accuracy, variance, and skewness. As a consequence, we are able to investigate a broad range of distributions (Figure 2A). For each of these populations, we repeatedly and randomly sample groups of n decision makers (i.e., individuals characterized by a given accuracy level; n = 1, 3, 5, and 9). Each of these n decision makers in a group faces the same m binary cases (m = 10, 25, and 100), and they independently judge the m cases, based on their accuracy level. Next, we pool the decisions of the n decision makers for each of the m cases, using a majority rule: if the majority of decision makers classifies the case correctly (incorrectly), the collective decision is categorized as correct (incorrect). We then calculate the group performance by summarizing the majority outcomes across the m cases (i.e., proportion correct decisions). For each unique combination of group size n, cases m, and accuracy distribution we sample 10,000 groups. Finally, for each unique treatment

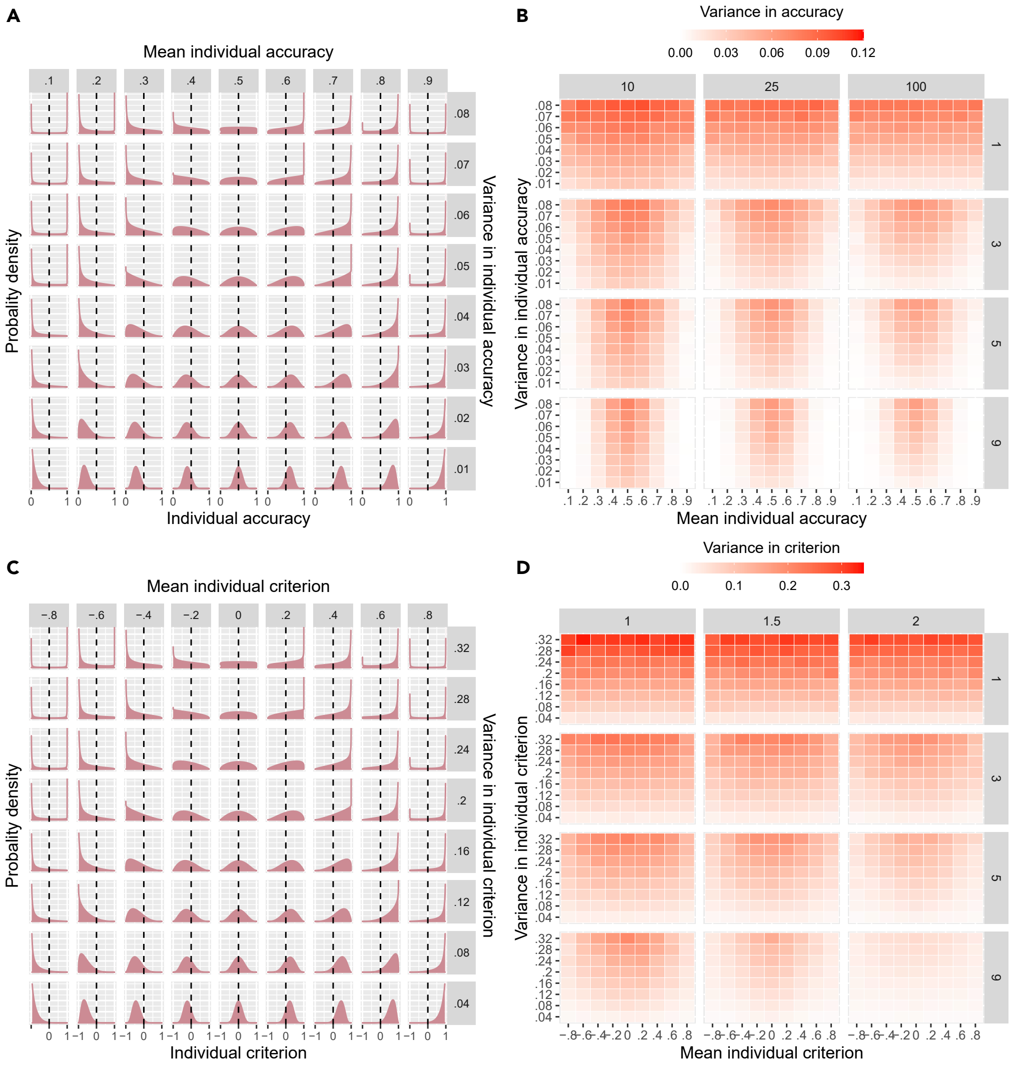


Figure 2. Numerical simulations: Pooling independent decisions reduces variation in accuracy and response bias in heterogeneous groups

(A) For the numerical simulations, we sampled decision makers from a wide range of populations with different performance distributions (x axis: individual accuracy; y axis: probability density). We generated those populations by systematically varying the mean (values at the top) and variance (values on the right) of the beta distribution. Dashed vertical lines indicate chance level (i.e., accuracy of 0.5).

(B) The variance in accuracy between individuals/groups for different-sized groups (n = 1, 3, 5, and 9; subpanel rows) making m = 10, 25, and 100 decisions (subpanel columns). Within each subpanel, the tiles correspond to the populations (i.e., accuracy distributions) of the respective tiles (i.e., mean-variance combination) in (A). The middle column within each subpanel corresponds to rater populations performing on average at chance level (i.e., mean accuracy = 0.5), with tiles to the left (right) of that column indicating populations performing on average below (above) chance. Darker color indicates increasing variance in accuracy between individuals or groups employing a majority rule. For each unique combination of group size n, decision cases m, and accuracy distribution, the variance shown corresponds to the variance between 10,000 independently sampled groups. As can be seen, independent of the specific accuracy distribution and number of decision cases considered, we robustly observe that increasing group size reduces the variance in accuracy between groups. This effect is smallest for populations at chance level and increases the closer the mean accuracy moves towards 1 or 0.

(C) For the simulations of response bias, we sampled decision makers from a wide range of populations of decision makers differing in their criterion value (x axis: criterion parameter; y axis: probability density). We generated those populations by varying the mean (values at the top) and variance (values on the right) of the beta distribution, where we transformed the beta range from [0,1] to [−1,1] to achieve a broad symmetrical range of criterion values. Dashed vertical lines indicate no response bias (i.e., criterion value = 0), raters increasingly to the left (right) of that line put increasingly

Figure 2. *Continued*

more weight on accuracy in state 1 (state 2). Note that we assume that all individuals within any given population have the same discrimination ability.

(D) The variance in response bias between individuals/groups for different-sized groups (n = 1, 3, 5, and 9; subpanel rows) and three different discrimination abilities 1, 1.5, and 2 (subpanel columns). Within each subpanel, the tiles correspond to populations (i.e., criterion distributions) of the respective tiles (i.e., mean–variance combination) in (C). Darker color indicates increasing variance in response bias between individuals, or groups employing a majority rule. For each unique combination of group size n, discrimination ability d', and response bias c, we independently sampled 10,000 groups and calculated the variance in response bias between these 10,000 groups. Independent of the specific distribution and discrimination ability in the population, we find that increasing group size robustly reduces the variation in response bias between groups.

combination, we calculate the variance in accuracy among the corresponding 10,000 groups as a measure of variation in accuracy between decision-making agents.

Figure 2B shows the results. Within each subpanel, the tiles correspond to the populations (i.e., accuracy distributions) of the respective tiles in Figure 2A. Moving from left to right within each subpanel corresponds to an increase in the mean accuracy of populations, and moving from bottom to top corresponds to an increase in the variance in individual accuracy. The middle column within each subpanel corresponds to populations performing at chance level (i.e., mean accuracy = 0.5), and tiles to the left (right) indicate populations performing below (above) chance. As can be seen from Figure 2B, independent of the specific accuracy distribution, the variation in accuracy between individuals is larger than that between groups of decision makers and increasing group size reduces the variation in performance between groups (i.e., different decision-making agents). This effect is smallest for populations at chance level and increases the closer the mean accuracy moves towards 1 or 0. The reduction in variation in accuracy between groups with increasing group size is present in almost all scenarios, suggesting this is a robust effect. Figure S1 shows that, as expected, the reduction of variation in accuracy between groups increases even further with increasingly larger groups. Figure S2 shows that, as expected, (1) when mean individual accuracy is above 0.5, the decreasing variation in accuracy between groups is associated with increasing mean group accuracy, whereas (2) when mean individual accuracy is below 0.5, the decreasing variation in accuracy between groups is associated with decreasing mean group accuracy.

Numerical simulations: Pooling decisions reduces variation in response bias in heterogeneous groups

We next focus on how the pooling of decisions affects variation in response bias, which is relevant whenever the world can be in one of two states (e.g., cancer present or absent; defendant guilty or not) and decision makers can thus commit two types of errors (false-negative versus false-positive decisions) (Macmillan and Creelman, 2005; Marshall et al., 2019; Sorkin and Dai, 1994; Sorkin et al., 2001). In order to reduce the number of possible scenarios to a feasible set, we make two simplifying assumptions: first, both states of the world appear equally often; second, decision makers within a population have the same ability to discriminate between signal present and signal absent cases. Following the simplest signal detection approach for such a setup (equal-variance Gaussian model) (Macmillan and Creelman, 1990, 2005), we use d' ($d' = z(HR) - z(FPR)$) as a measure of discrimination ability (which we vary from 1 to 1.5 to 2 between populations) and the criterion value c ($c = -(z(HR) + z(FPR))/2$) as a measure of response bias, where $z(.)$ is the inverse distribution function of the normal distribution, yielding the z-scores for the hit rate (HR; also known as sensitivity) and the false-positive rate (FPR; 1 – specificity), respectively (Macmillan and Creelman, 1990, 2005). We use the criterion value c as a widely used measure of response bias (Macmillan and Creelman, 2005). Analogous to the accuracy analysis above, we generate a large range of populations varying in their distribution of response biases. We do this by systematically varying the mean and the variance of the associated criterion distribution, allowing the criterion of individuals to range between −1 and +1 (Figure 2C). For each of these populations, we repeatedly and randomly sample groups of n decision makers (i.e., individuals characterized by a particular criterion value; n = 1, 3, 5, and 9), holding constant the number of cases to be decided upon (1,000). For each of the simulated groups that pool decisions with a majority rule, we then calculate the accuracy (i.e., proportion correct) separately for state 1 (truly positive cases: sensitivity) and state 2 (truly negative cases: specificity) and use both values to calculate the implied criterion value (see STAR methods). We report sensitivity and specificity separately because the trade-off between the two types of errors is important in many real-world settings (e.g., medical diagnostics or forensics, which we investigate in the empirical analysis below). For each combination of group size n, discrimination ability d', and criterion distribution, we again independently sample 10,000 groups and calculate the variance in criterion among the corresponding 10,000 groups as a measure of variation in response bias.

Figure 2D shows the results of this analysis with darker color indicating higher variance in response bias between groups. Within each subpanel, the tiles correspond to the populations of the respective tiles in Figure 2C. Independent of the specific criterion distribution and discrimination ability, increasing group size reduces the variation in response bias between decision-making agents. That is, across a wide range of populations that differ in how much individuals vary in their response biases, we find that the pooling of decisions substantially reduces variation in response bias between agents. This reduction is observed in almost all scenarios, suggesting this is a robust effect. Figure S3 shows that, as expected, increasing group size also leads to a reduction in variance in sensitivity and specificity between groups.

For simplicity, so far we assumed that the decisions of individuals in the same group were statistically independent of one another. Figure S4 shows how variation in accuracy and response bias develops as the correlation between decisions increases (see STAR methods for a description of how the correlations were implemented). When increasing the correlation between decisions, the key result still holds: increasing group size reduces variation in accuracy and criterion.

Empirical analysis: Pooling decisions reduces variation in accuracy and response bias in real-world contexts

Our theoretical and simulation results suggest that a decision-making system that pools decisions represents a simple and effective approach to reduce outcome variation between decision-making agents in binary classification tasks. To corroborate this prediction in real-world contexts, we next analyze several published datasets from five domains: (1) a breast cancer dataset, comprising 15,655 diagnoses by 101 radiologists based on 155 mammograms (Carney et al., 2012); (2) a skin cancer dataset, comprising 4,320 diagnoses by 40 dermatologists based on 108 dermoscopic images of skin lesions (Argenziano et al., 2003); (3) a fingerprint recognition dataset, comprising 1,584 evaluations by 36 professional fingerprint examiners of 44 fingerprint pairs (Tangen et al., 2020), (4) a geopolitical forecasting dataset from the Good Judgment Project, containing 8,258 forecasts by 89 forecasters on 94 geopolitical events (Ungar et al., 2012); and (5) a general knowledge dataset, containing 99,000 responses by 99 individuals to 1,000 questions (here: which of two cities is larger) (Yu et al., 2015). All datasets are described in detail in the STAR methods. For the medical datasets, the patient's actual health state (i.e., cancer present versus absent) was known from follow-up research; for the forecasting dataset, the correctness of forecasts was determined by follow-up research. We use all datasets to investigate the consequences of pooling decisions for variation in accuracy. Figures 3A–3C, 4A, 4B, and S5A–S5C show that individuals vary substantially in their accuracy levels in all five datasets. Response bias could, however, only be studied in three of the datasets (breast cancer, skin cancer, and finger recognition) as there is no clear distinction between false-negative and false-positive errors in the forecasting and the general knowledge datasets. In the three former datasets, individuals also vary substantially in their response bias (Figures 5A–5C).

To investigate how variation in accuracy and response bias change with group size we randomly and repeatedly sample (without replacement) two groups of size n (1, 3, 5, 7, and 9) within each dataset. For each draw of two groups, we calculate the performance of each group across all cases under the majority rule and—as a measure of variation in performance between groups—the absolute difference in performance between both groups. We calculate performance in terms of sensitivity, specificity, and d' for breast and skin cancer, and fingerprint recognition, and in terms of overall accuracy for geopolitical forecasting and general knowledge (see STAR methods for calculation of d'). For breast and skin cancer, and fingerprint recognition, we also calculate the criterion value c and—as a measure of variation in response bias between groups—the absolute difference in the criterion between both groups. Within each dataset we repeat this procedure 1,000 times for each group size (i.e., we independently sample two groups of a given size) and determine the mean values of all the above measures across these 1,000 draws. This approach yields the mean pairwise absolute difference (MPAD) between two groups (i.e., the expected difference between two randomly sampled groups). We chose this pairwise measure of variation as it compares the outcomes of groups consisting of different individuals. Using the variance (a more common measure of variation and the measure we used in the numerical simulations) would imply comparing groups consisting of partly overlapping members, which could confound our results on the relationship between group size and outcome variation.

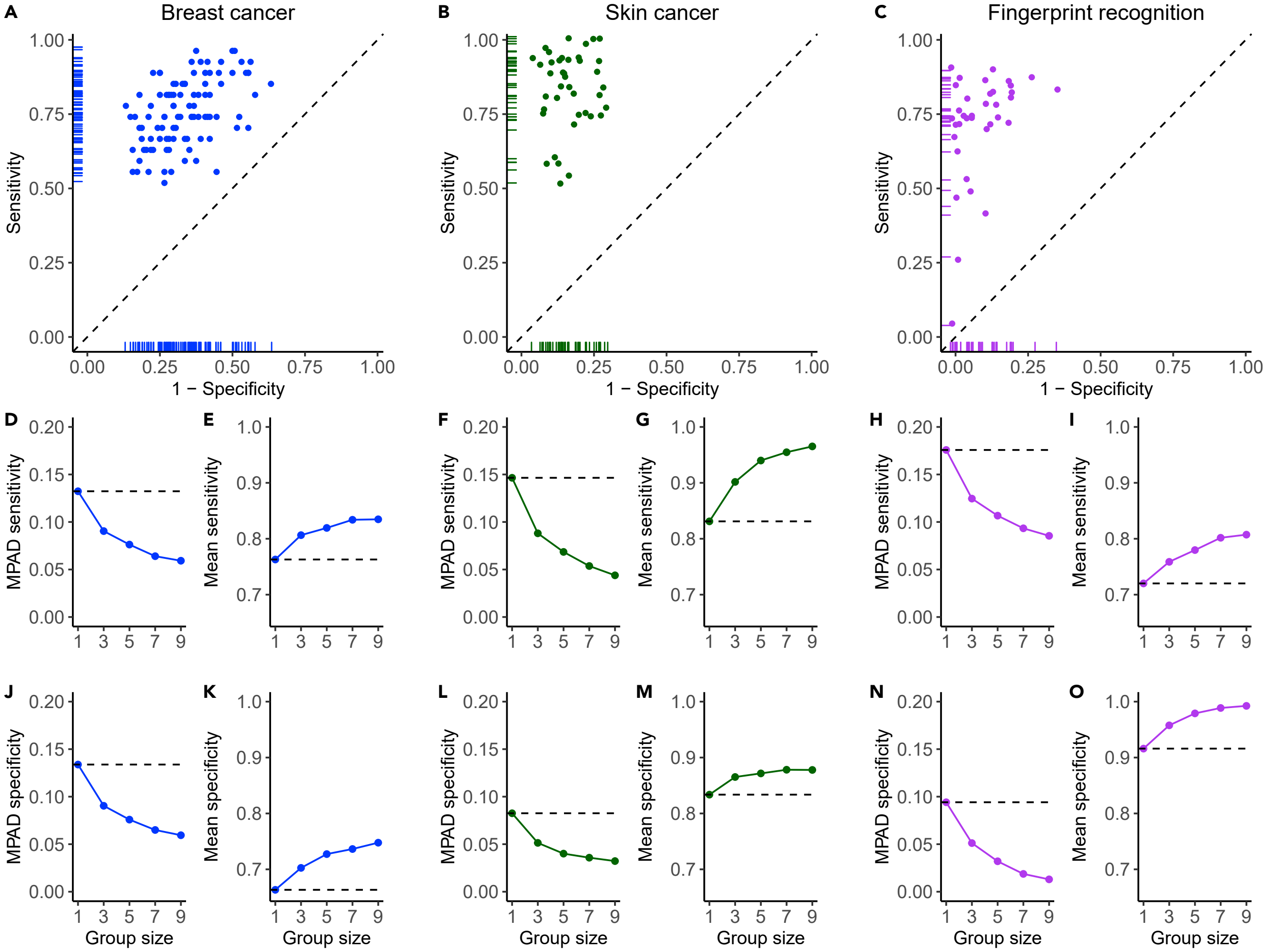


Figure 3. Pooling independent decisions reduces variation in sensitivity and specificity in breast and skin cancer detection and in fingerprint recognition

(A–C). True- and false-positive rates (i.e., sensitivity and 1 − specificity, respectively) of each rater in the three datasets. Each dot corresponds to one rater. The dashed diagonal corresponds to those points that can be achieved by a random classifier; dots above (below) the diagonal indicate performance above (below) chance. In all three domains, raters differ substantially in both accuracy components (i.e., sensitivity and specificity).

(D–O) For each dataset and each group size, we repeatedly and randomly sampled two groups and calculated the mean pairwise absolute differences (MPAD) in sensitivity and specificity. As predicted, in all three datasets, relative to the baseline levels of variation between individual decision makers (dashed lines), substantial reductions in variation (i.e., MPAD) in (D, F, H) sensitivity and (J, L, N) specificity are achieved by pooling independent decisions; as expected, these reductions increase with increasing group size. These reductions in variation are accompanied by an increase in mean (E, G, I) sensitivity and (K, M, O) specificity.

Figures 3D–3O, 4C–4F, and 5D–5I show the results of this analysis. As predicted, the mean pairwise absolute difference (MPAD) in sensitivity (Figures 3D, 3F, and 3H) and specificity (Figures 3J, 3L, and 3N) decreases with increasing group size in the breast and skin cancer, and fingerprint recognition datasets. This reduction is substantial in all scenarios, and strong reductions already occur at relatively small group sizes. For example, the MPAD in sensitivity between two randomly selected single experts is 0.13 for breast cancer, 0.13 for skin cancer, and 0.17 for fingerprint recognition; pooling the decisions of five experts reduces these values to 0.06, 0.06, and 0.10, which amounts to relative reductions of 52%, 54%, and 41%, respectively. Of importance, and as reported before (Kurvers et al., 2015; Tangen et al., 2020; Wolf et al., 2015), mean sensitivity (Figures 3E, 3G, and 3I) and specificity (Figures 3K, 3M, and 3O) increase with increasing group size in all three datasets. Figures S5D–S5I show that the same patterns emerge for the performance measure *d*′. Similarly, in the geopolitical forecasting and general knowledge dataset, increasing group size substantially reduces the MPAD in accuracy between groups (Figures 4C and 4D) while increasing mean accuracy (Figures 4E and 4F). Figure S6 shows that similar results emerge when

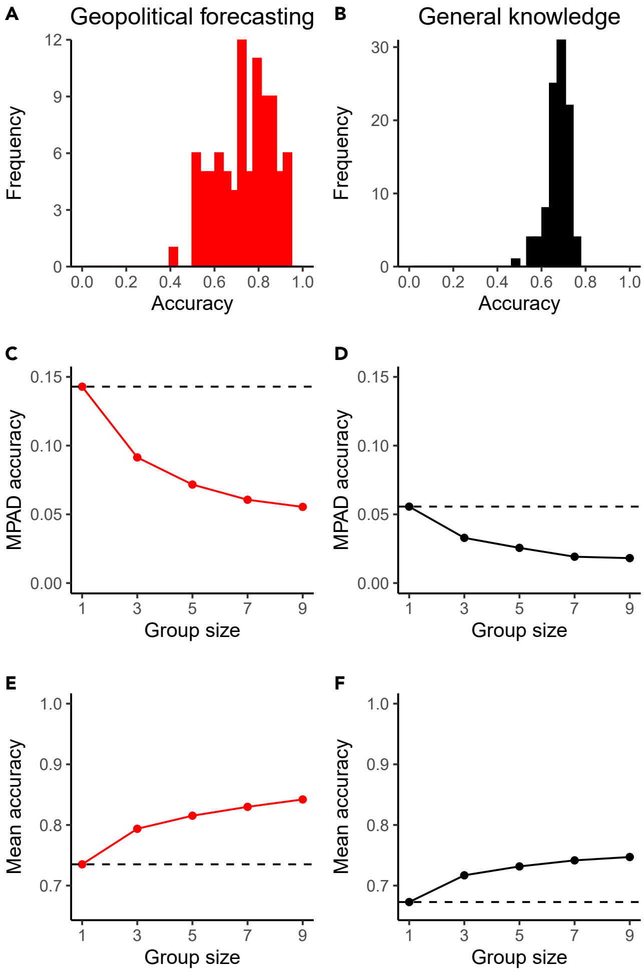


Figure 4. Pooling independent decisions reduces variation in geopolitical forecasting and in a general knowledge task

(A and B) Frequency distribution of the accuracy levels of raters in the (A) geopolitical forecasting and (B) general knowledge dataset, showing that raters in both domains differ substantially in accuracy levels.

(C and D) For each dataset and each group size (n = 1, 3, 5, 7, and 9), we repeatedly and randomly sampled two groups and—as a measure of variation between groups—calculated the mean pairwise absolute differences (MPAD) in accuracy. As predicted, in both datasets, relative to the baseline level of variation between individual decision makers (dashed lines), pooling independent decisions substantially reduces variation (i.e., MPAD) in accuracy; as expected, these reductions increase with larger group sizes.

(E and F) These reductions in variation are accompanied by an increase in mean accuracy in both datasets.

we use the original continuous subjective probability scale in the forecasting dataset (rather than the binary yes/no scale).

As predicted, the MPAD in criterion value (i.e., response bias) also decreases with increasing group size in the breast and skin cancer, and fingerprint recognition datasets (Figures 5D–5F). Again, this reduction is substantial in all three datasets and even with relatively small group sizes. The mean criterion value either remains the same or slightly increases or decreases with group size (Figures 5G–5I). Thus, as predicted, pooling independent decisions robustly reduces differences in response bias and accuracy between decision-making agents.

Empirical analysis: Pooling decisions reduces variation at the case level

Thus far, we have focused on differences between decision-making agents across a large number of cases. Next, we ask whether and to what extent collective decision-making systems based on the pooling of

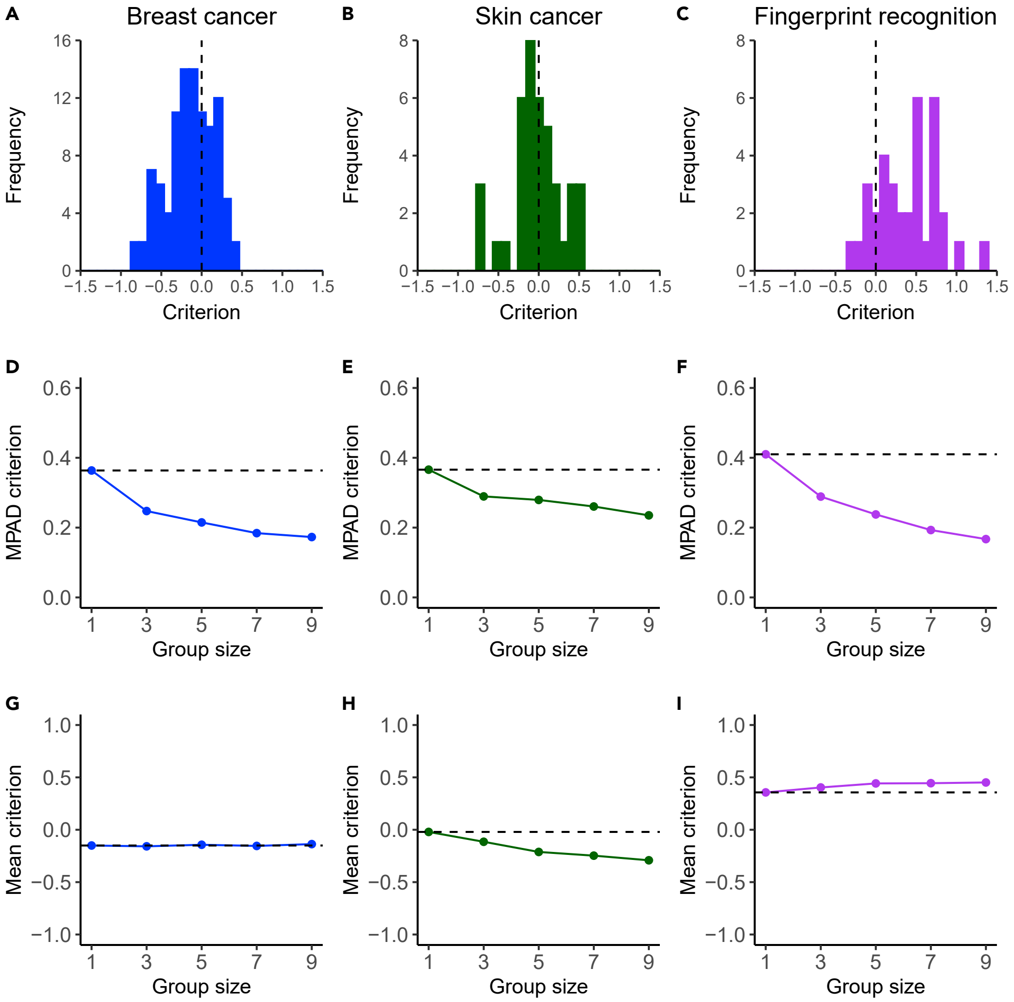


Figure 5. Pooling independent decisions reduces variation in response bias in breast and skin cancer detection and in fingerprint recognition

(A–C) Frequency distribution of the criterion values of raters in the three datasets, showing that raters in each of these domains differ substantially in criterion values. Dashed vertical lines show the neutral criterion value.

(D–I) For each dataset and each group size, we repeatedly and randomly sampled two groups and calculated the mean pairwise absolute differences (MPAD) in the criterion (i.e., response bias). (D–F) As predicted, in all three datasets, pooling independent decisions substantially reduces variation (i.e., MPAD) in the decision criterion. The mean criterion either (G) remains unchanged, (H) decreases slightly, or (I) increases slightly.

independent decisions can also reduce variation on the level of the individual case, be it a mammogram, skin lesion, fingerprint set, forecasting question, or general knowledge question. To this end, we randomly and repeatedly sample two individuals and determine whether or not both individuals give the same response to each individual case in each dataset. This is repeated 2,500 times per case. We then calculate the average frequency of agreement per case. We use the same procedure for groups: for each case within each dataset, we randomly and repeatedly sample two groups of *n* individuals (5 and 9) and determine whether or not these two groups arrive at the same response under the majority rule. This is repeated 2,500 times per unique combination of case and group size; we then calculate the average frequency of agreement per case per group size.

Figure 6 shows the results of this analysis. In all datasets, there are many cases with substantial between-individual disagreement. For example, in the breast cancer dataset, 71 of 155 cases have an agreement level below 0.6, implying an at least 40% chance that two randomly sampled radiologists will disagree.

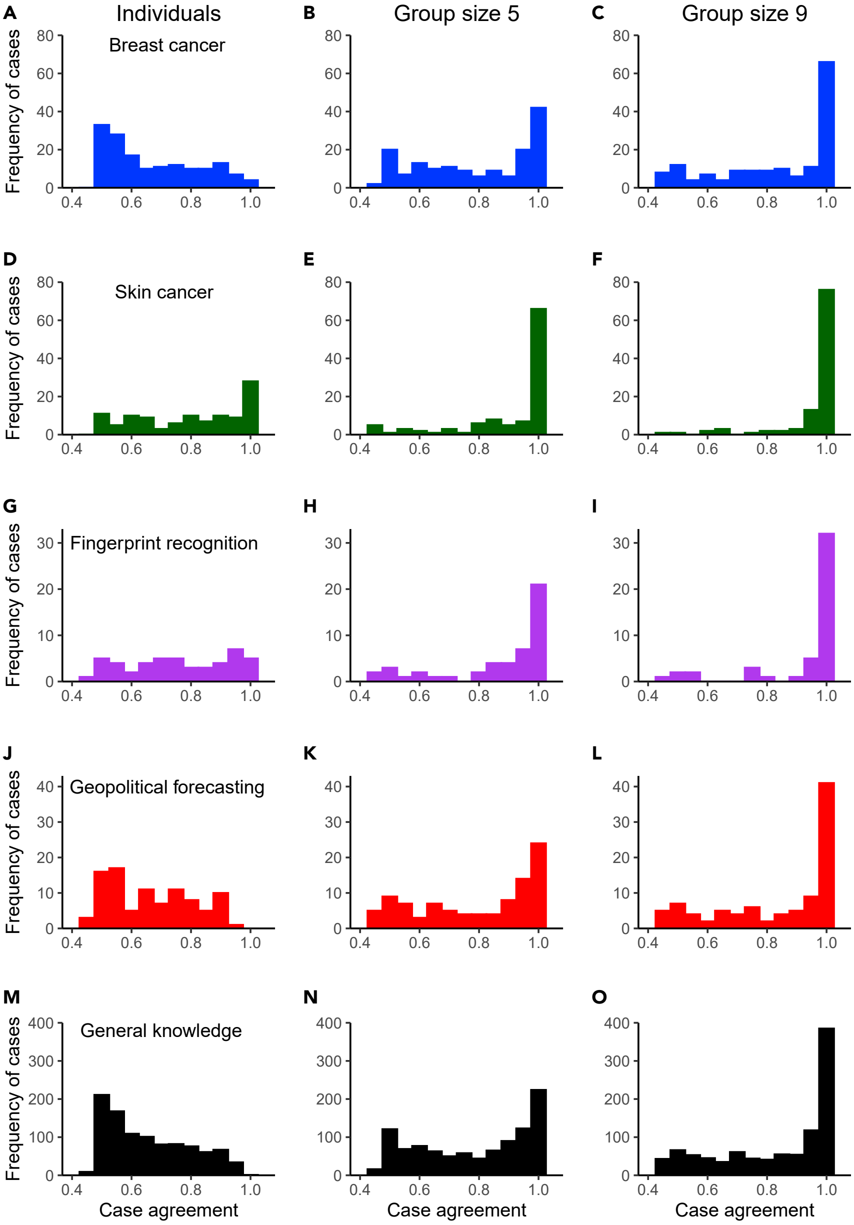


Figure 6. Pooling independent decisions reduces variation at the case level in all five datasets

Histograms show for each case the proportion of comparisons where (1) two randomly selected individuals, (2) two groups of five randomly selected individuals, and (3) two groups of nine randomly selected individuals agree, for (A–C) breast cancer, (D–F) skin cancer, (G–I) fingerprint recognition, (J–L) geopolitical forecasting, and (M–O) a general knowledge task. In all five datasets, there is a substantial frequency of cases where two randomly sampled individuals disagree. Relative to this baseline level, pooling independent decisions systematically decreases the number of situations where decision-making agents disagree and increases the number of situations where they agree (i.e., the distributions shift to the right); this effect increases with increasing group size.

In all datasets, this distribution shifts toward more agreement with increasing group size. That is, the pooling of decisions systematically decreases the number of situations where decision-making agents disagree and increases the number of situations where they agree. Pooling decisions thus also reduces outcome variation between decision-making systems at the level of individual cases.

DISCUSSION

Collective decision-making systems based on the pooling of independent decisions can be a relatively simple and effective approach to boost decision accuracy (Grofman et al., 1983; Hastie and Kameda, 2005; Herzog et al., 2019; Mannes et al., 2014; Surowiecki, 2004). But it can do even more. Combining a theoretical and data-driven approach, we analyzed a previously mostly neglected performance dimension of collective and individual decision-making systems: the variation in outcomes between decision-making agents. Our theoretical and empirical data-driven analyses converge to the same conclusion: pooling independent judgments is a powerful method to reduce outcome variation between decision-making agents. As has been widely demonstrated, individual experts differ substantially in terms of both accuracy and response bias (Burgman, 2016; Deneef and Kent, 1993; Koran, 1975; Kurvers et al., 2016; Mellers et al., 2015; O'Sullivan and Ekman, 2004; Wolf et al., 2015) (see also Figures 3A–3C and 5A–5C). Our results suggest that this, in many contexts, undesirable variation can be substantially reduced by combining their independent decisions.

A key implication of our findings is that pooling independent opinions will often be more reliable and predictable than betting on an individual expert, thus promoting actual and perceived fairness (Tyler et al., 1996; Wood et al., 2020). In judicial decision making, for example, the principle of equal treatment under the law suffers when judges differ in their response bias (i.e., how much evidence they require to convict a suspect). Relatedly, people perceive outcomes to be fairer (and are more likely to comply) if they believe that the personal values and biases of professional decision makers did not enter the decision making process (Tyler, 2000). Beyond boosting fairness, decision-making systems that are more reliable and predictable—due to lower outcome variation—are typically also perceived as more trustworthy. Collective systems that pool individual decisions may be one tool to prevent the erosion of institutional trust.

Throughout, we have argued that low variation at the outcome level is beneficial. This, however, need not always be the case. Pooling independent decisions can in some cases decrease accuracy—for example, when average individual accuracy is below chance (Boland, 1989). Critically, pooling decisions in this scenario would also decrease outcome variation (see our statistical argument and numerical simulations), thereby increasing the reliability—and thus potentially trust in the wrong majority decision. Relatedly, whenever the biases of collective systems are large (e.g., the judged probabilities of a disease being present is generally too high in a population of experts), then a system that maintains variation may be preferable to one with reduced variation because the former is more likely to make correct decisions (e.g., a healthy case is more likely to be categorized as healthy because, with higher variance, there is a larger chance that the judged probability will be below the diagnostic cutoff) (see also Friedman, 1997).

Response biases play a crucial role in decision making in a wide range of domains, including medical, judicial, and political decision making (DeKay, 1996; Swets et al., 2000; Wolf et al., 2013). A key normative question here is how much evidence should be required to classify a mammogram as malignant, or to convict a defendant (DeKay, 1996; Deneef and Kent, 1993; Swets, 1992). Individual decision makers differ with respect to the decision threshold (i.e., response bias), which, in turn, can contribute to differences in decisions outcomes among decision makers. Collective systems based on pooling independent decisions can be a powerful corrective to such unwanted variation. In many contexts it may be both unclear what exactly a normatively defensible threshold would be and that threshold might differ among the affected population of people. For example, patients undergoing medical treatment often differ in their preferences (O'Connor et al., 2003; Schwartz et al., 2000), and taking these individual preferences into account can be important when providing treatment recommendations. Pooling independent judgments can also be useful for such settings with heterogeneous preferences because a lower variation in response biases among the decision-making agents will reduce the gap between the desired and enacted response bias. Although we have focused on the majority rule in our analyses, independent decisions can also be aggregated using more flexible quorum thresholds that allow decisions to be fine-tuned under any error cost scheme (Marshall et al., 2019; Wolf et al., 2013; Kurvers et al., 2015) illustrates this approach for the context of skin cancer diagnostics.

Many previous studies have investigated how the variation (or diversity) between individuals in a group affects its collective accuracy (Hong and Page, 2004). It has been shown that even mild social influence can undermine collective accuracy by decreasing variation between individuals (Lorenz et al., 2011). Here, in contrast, we focused on variation at the outcome level, that is, the variation between different groups of decision makers. Pitting past and present results side by side creates an interesting juxtaposition. Although previous studies have shown that collective decision-making systems often benefit from variation (diversity) at the level of the input (i.e., in the opinions of different individuals), these systems will also, at the same time, reduce the often undesirable outcome variation.

The importance of outcome variation in forecasts, judgments and decisions has been pointed out across the behavioral sciences (Kahneman et al., 2016; Litvinova et al., 2019; Stewart, 2001). For example, in the literature on optimal portfolio selection, the importance of taking into account both the mean rate and the variance (i.e., risk) of returns on securities is well established (Brealey et al., 2012); likewise, in the literature on forecasting (Hibon and Evgeniou, 2005; Lichtendahl and Winkler, 2020) and machine learning (Kuncheva, 2014), aggregation has been shown to reduce risk. In machine learning, classifier ensembles are known to reduce the variance of the answers, which increases accuracy (Kuncheva, 2003, 2014; Kuncheva and Whitaker, 2003). In animal behavior, the theory on risk-sensitive foraging is centered on the insight that—next to average foraging returns—variation in foraging returns over time is a key fitness determinant (McNamara and Houston, 1992). Our findings may stimulate new theoretical and empirical questions in these contexts.

The strength and beauty of studying binary classification tasks lies in the combination of broad applicability to real-world problems (from medical diagnostics to democratic processes) with the fact that the outcomes of binary decision processes are governed by fundamental statistical principles (Hastie and Kameda, 2005). Much previous research has focused on the Condorcet Jury Theorem (and variations thereof), which uses basic properties of the binomial distribution to understand how pooling decisions affects accuracy levels (Boland, 1989; Grofman et al., 1983; Marshall et al., 2019). As we have shown, the basic properties of the binomial distribution also cast light on the relationship between pooling decisions and variation in decision outcomes between agents. In a nutshell, the binomial distribution provides a direct link between accuracy levels and the similarity of decisions between agents (Kurvers et al., 2019). In combination with the effect of pooling decisions on accuracy levels, this turns out to be a key factor shaping the variation in outcomes between agents.

In recent decades, researchers across the behavioral sciences have carefully mapped out the relative advantages and disadvantages of individual and collective decision-making systems in terms of accuracy. Variation in outcomes between decision-making agents is another key benchmark of decision-making systems. As our results demonstrate, collective decision-making systems based on the pooling of independent decision are a powerful approach to reduce such variation, thus promoting reliability, predictability, fairness, and possibly even trust.

Limitations of the study

We have focused on pooling independent judgments; collective decisions can, of course, arise in various ways. One question for future research is whether other systems of collective decision making, such as discussion among interacting individuals, also achieve higher levels of between-group agreement. Directly interacting individuals can, under some circumstances, achieve higher accuracy than individual decision makers and/or systems based on pooling independent judgments (Navajas et al., 2018). However, little is known about the consequences for variation in outcomes between different interacting groups. Although pooling independent judgments is firmly rooted in statistical principles, the dynamics of interacting groups are governed by, for instance, social and conversational norms (e.g., which pieces of information are mentioned during a discussion and in what order [Stasser and Abele, 2020; Stasser and Titus, 1985]). For example, cognitive strategies, such as a confirmation strategy, may be amplified when individuals with similar biases (e.g., in terms of response bias) interact. In mock juries deliberation in groups leads to more leniency in sentencing, relative to an individual reaching a verdict by themselves (MacCoun and Kerr, 1988). Whether and how collective systems based on interacting groups affect variation in outcomes between groups is thus not straightforward and requires further theoretical and empirical scrutiny. Another issue closely related to our work concerns biases, for example, based on gender or race, in health care (FitzGerald and Hurst, 2017; Hall et al., 2015) and court rooms (Franklin, 2018; Spohn, 2000). Future work could investigate whether collective systems such as pooling or interacting groups could also be employed to reduce such biases.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
    - ○ Lead contact

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102740.

ACKNOWLEDGMENTS

We thank three anonymous referees for their constructive criticism. We thank Amy Buzby, Jose Cayere, and the American College of Radiology for their technical assistance in developing and supporting the implementation of the test sets; the expert radiologists Larry Bassett, Barbara Monsees, Ed Sickles, and Matthew Wallis; and the participating women, facilities, and radiologists for the data they provided. The BCSC investigators are listed at www.breastscreening.cancer.gov/. This work 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-09-271-01, and SIRSG-06-290-04); by the Breast Cancer Stamp Fund; and by the National Cancer Institute BCSC (Grant HHSN261201100031C). The collection of cancer and vital status data used in this study was supported by several state public health departments and cancer registries throughout the United States. A full description of these sources is provided at www.breastscreening.cancer.gov/work/acknowledgement.html. This work was further funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2002/1 ''Science of Intelligence''—project number 390523135. We thank Susannah Goss for editing the manuscript.

AUTHOR CONTRIBUTIONS

R.H.J.M.K., S.M.H., R.H., J.K., and M.W. conceived the original idea. M.W. formulated the general statistical argument. R.H.J.M.K. performed the numerical simulations and the simulations of the experimental data with input from S.M.H. R.H.J.M.K. and M.W. wrote the manuscript with substantial input from all other authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 16, 2020
Revised: May 3, 2021
Accepted: June 15, 2021
Published: July 23, 2021

REFERENCES

Argenziano, G., Soyer, H.P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., De Rosa, G., Ferrara, G., et al. (2003). Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. J. Am. Acad. Dermatol. *48*, 679–693. https://doi.org/10.1067/mjd.2003.281.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science *329*, 1081–1085. https://doi.org/10.1126/science.1185718.

Bang, D., and Frith, C.D. (2017). Making better decisions in groups. R. Soc. Open Sci. *4*, 170193. https://doi.org/10.1098/rsos.170193.

Boland, P.J. (1989). Majority systems and the Condorcet jury Theorem. The Statistician *38*, 181. https://doi.org/10.2307/2348873.

Brealey, R.A., Myers, S.C., Allen, F., and Mohanty, P. (2012). Principles of Corporate Finance (Tata McGraw-Hill Education).

Burgman, M.A. (2016). Trusting Judgements: How to Get the Best Out of Experts (Cambridge University Press).

Carney, P.A., Bogart, T.A., Geller, B.M., Haneuse, S., Kerlikowske, K., Buist, D.S., Smith, R., Rosenberg, R., Yankaskas, B.C., Onega, T., and

Miglioretti, D.L. (2012). Association between time spent interpreting, level of confidence, and accuracy of screening mammography. Am. J. Roentgenol. *198*, 970–978. https://doi.org/10.2214/AJR.11.6988.

Clément, R.J.G., Krause, S., von Engelhardt, N., Faria, J.J., Krause, J., and Kurvers, R.H.J.M. (2013). Collective cognition in humans: groups outperform their best members in a sentence reconstruction task. PLoS One *8*, e77943. https://doi.org/10.1371/journal.pone.0077943.

Conradt, L., and List, C. (2009). Group decisions in humans and animals: a survey. Philos. Trans. R Soc. B *364*, 719–742. https://doi.org/10.1098/rstb.2008.0276.

DeKay, M.L. (1996). The difference between Blackstone-like error ratios and probabilistic standards of proof. Law Soc. Inq. *21*, 95–132. https://doi.org/10.1111/j.1747-4469.1996.tb00013.x.

Deneef, P., and Kent, D.L. (1993). Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis. Med. Decis. Mak. *13*, 126–132. https://doi.org/10.1177/0272989X9301300206.

El Zein, M., Bahrami, B., and Hertwig, R. (2019). Shared responsibility in collective decisions. Nat. Hum. Behav. *3*, 554–559. https://doi.org/10.1038/s41562-019-0596-4.

FitzGerald, C., and Hurst, S. (2017). Implicit bias in healthcare professionals: a systematic review. BMC Med. Ethics *18*, 19. https://doi.org/10.1186/s12910-017-0179-8.

Franklin, T.W. (2018). The state of race and punishment in America: is justice really blind? J. Crim. Justice *59*, 18–28. https://doi.org/10.1016/j.jcrimjus.2017.05.011.

Friedman, J.H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data Min. Knowl. Discov. *1*, 55–77. https://doi.org/10.1023/A:1009778005914.

Geller, B.M., Bogart, A., Carney, P.A., Sickles, E.A., Smith, R., Monsees, B., Bassett, L.W., Buist, D.M., Kerlikowske, K., and Onega, T. (2014). Educational interventions to improve screening mammography interpretation: a randomized controlled trial. Am. J. Roentgenol. *202*, W586–W596. https://doi.org/10.2214/AJR.13.11147.

Grofman, B., Owen, G., and Feld, S.L. (1983). Thirteen theorems in search of the truth. Theor. Decis. *15*, 261–278. https://doi.org/10.1007/BF00125672.

Gully, S.M., Incalcaterra, K.A., Joshi, A., and Beaubien, J.M. (2002). A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. J. Appl. Psychol. *87*, 819–832. https://doi.org/10.1037/0021-9010.87.5.819.

Hall, W.J., Chapman, M.V., Lee, K.M., Merino, Y.M., Thomas, T.W., Payne, B.K., Eng, E., Day, S.H., and Coyne-Beasley, T. (2015). Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am. J. Public Health *105*, e60–e76. https://doi.org/10.2105/AJPH.2015.302903.

Hammond, K.R. (1996). Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice (Oxford University Press on Demand).

Hastie, R., and Kameda, T. (2005). The robust beauty of majority rules in group decisions. Psychol. Rev. *112*, 494–508. https://doi.org/10.1037/0033-295x.112.2.494.

Herzog, S.M., Litvinova, A., Yahosseini, K.S., Novaes Tump, A., and Kurvers, R.H.J.M. (2019). The ecological rationality of the wisdom of crowds. In Taming Uncertainty, Hertwig., Pleskac., and Pachur., eds. (MIT Press).

Hibon, M., and Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. Int. J. Forecast. *21*, 15–24. https://doi.org/10.1016/j.ijforecast.2004.05.002.

Hong, L., and Page, S.E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc. Natl. Acad. Sci. U S A *101*, 16385–16389. https://doi.org/10.1073/pnas.0403723101.

Kahneman, D., Rosenfield, A., Gandhi, L., and Blaser, T. (2016). Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision-Making (HBR), pp. 38–46.

Kerr, N.L., and Tindale, R.S. (2004). Group performance and decision making. Annu. Rev. Psychol. *55*, 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009.

Koran, L.M. (1975). The reliability of clinical methods, data and judgments. N. Engl. J. Med. *293*, 642–646. https://doi.org/10.1056/NEJM197509252931307.

Koriat, A. (2012). When are two heads better than one and why? Science *336*, 360–362. https://doi.org/10.1126/science.1216549.

Kuncheva, L.I. (2003). That Elusive Diversity in Classifier Ensembles (Springer), pp. 1126–1138.

Kuncheva, L.I. (2014). Combining Pattern Classifiers: Methods and Algorithms (John Wiley & Sons).

Kuncheva, L.I., and Whitaker, C.J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. *51*, 181–207. https://doi.org/10.1023/a:1022859003006.

Kurvers, R.H., Herzog, S.M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., Zalaudek, I., Carney, P., and Wolf, M. (2019). How to detect high-performing individuals and groups: decision similarity predicts accuracy. Sci. Adv. *5*, eaaw9011. https://doi.org/10.1126/sciadv.aaw9011.

Kurvers, R.H., Krause, J., Argenziano, G., Zalaudek, I., and Wolf, M. (2015). Detection accuracy of collective intelligence assessments for skin cancer diagnosis. JAMA Dermatol. *151*, 1346–1353. https://doi.org/10.1001/jamadermatol.2015.3149.

Kurvers, R.H.J.M., Herzog, S.M., Hertwig, R., Krause, J., Carney, P.A., Bogart, A., Argenziano, G., Zalaudek, I., and Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. Proc. Natl. Acad. Sci. U S A *113*, 8777–8782. https://doi.org/10.1073/pnas.1601827113.

Lichtendahl, K.C., Jr., and Winkler, R.L. (2020). Why do some combinations perform better than others? Int. J. Forecast. *36*, 142–149. https://doi.org/10.1016/j.ijforecast.2019.03.027.

Litvinova, A., Kurvers, R.H., Hertwig, R., and Herzog, S.M. (2019). When Experts Make Inconsistent Decisions. https://osf.io/e7nk6/.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. Proc. Natl. Acad. Sci. U S A *108*, 9020–9025. https://doi.org/10.1073/pnas.1008636108.

MacCoun, R.J., and Kerr, N.L. (1988). Asymmetric influence in mock jury deliberation: jurors' bias for leniency. J. Pers Soc. Psychol. *54*, 21. https://doi.org/10.1037/0022-3514.54.1.21.

Macmillan, N.A., and Creelman, C.D. (1990). Response bias: characteristics of detection theory, threshold theory, and" nonparametric" indexes. Psychol. Bull. *107*, 401. https://doi.org/10.1037/0033-2909.107.3.401.

Macmillan, N.A., and Creelman, C.D. (2005). Detection Theory: A User's Guide, Second Edition (Lawrence Erlbaum Associates).

Mannes, A.E., Soll, J.B., and Larrick, R.P. (2014). The wisdom of select crowds. J. Pers Soc. Psychol. *107*, 276. https://doi.org/10.1037/a0036677.

Marshall, J.A., Kurvers, R.H., Krause, J., and Wolf, M. (2019). Quorums enable optimal pooling of independent judgements in biological systems. Elife *8*, e40368. https://doi.org/10.7554/eLife.40368.

McNamara, J.M., and Houston, A.I. (1992). Risk-sensitive foraging: a review of the theory. Bull. Math. Biol. *54*, 355–378. https://doi.org/10.1007/BF02464838.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., and Horowitz, M. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspect. Psychol. Sci. *10*, 267–281. https://doi.org/10.1177/1745691615577794.

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., and Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nat. Hum. Behav. *2*, 126–132. https://doi.org/10.1038/s41562-017-0273-4.

O'Connor, A.M., Légaré, F., and Stacey, D. (2003). Risk communication in practice: the contribution of decision aids. BMJ *327*, 736–740. https://doi.org/10.1136/bmj.327.7417.736.

O'Sullivan, M., and Ekman, P. (2004). 12 the Wizards of Deception Detection. The Detection of Deception in Forensic Contexts (Cambridge University Press), p. 269.

R-Core-Team. (2021). R: A Language and Environment for Statistical Computing.

Schwartz, L.M., Woloshin, S., Sox, H.C., Fischhoff, B., and Welch, H.G. (2000). US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. BMJ *320*, 1635–1640. https://doi.org/10.1136/bmj.320.7250.1635.

Sorkin, R.D., and Dai, H. (1994). Signal detection analysis of the ideal group. Organ. Behav. Hum. Decis. Process. *60*, 1–13. https://doi.org/10.1006/obhd.1994.1072.

Sorkin, R.D., Hays, C.J., and West, R. (2001). Signal-detection analysis of group decision making. Psychol. Rev. *108*, 183–203. https://doi.org/10.1037//0033-295x.108.1.183.

Spohn, C. (2000). Thirty Years of Sentencing Reform: The Quest for a Racially Neutral Sentencing Process (National Institute of Justice).

Stasser, G., and Abele, S. (2020). Collective choice, collaboration, and communication. Annu. Rev. Psychol. *71*, 589–612. https://doi.org/10.1146/annurev-psych-010418-103211.

Stasser, G., and Titus, W. (1985). Pooling of unshared information in group decision making: biased information sampling during discussion. J. Pers. Soc. Psychol. *48*, 1467. https://doi.org/10.1037/0022-3514.48.6.1467.

Stewart, T.R. (2001). Improving reliability of judgmental forecasts. In Principles of Forecasting, Armstrong., ed. (Springer), pp. 81–106.

Surowiecki, J. (2004). 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).

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. Am. Psychol. *47*, 522–532. https://doi.org/10.1037/0003-066X.47.4.522.

Swets, J.A., Dawes, R.M., and Monahan, J. (2000). Psychological science can improve diagnostic decisions. Psychol. Sci. Public Interest *1*, 1–26. https://doi.org/10.1111/1529-1006.001.

Tangen, J.M., Kent, K., and Searston, R.A. (2020). Collective intelligence in fingerprint analysis. Cogn. Res. Princ Implic. *5*. https://doi.org/10.1186/s41235-020-00223.

Tyler, T.R. (2000). Social justice: outcome and procedure. Int. J. Psychol. *35*, 117–125. https://doi.org/10.1080/002075900399411.

Tyler, T., Degoey, P., and Smith, H. (1996). Understanding why the justice of group procedures matters: a test of the psychological dynamics of the group-value model. J. Pers. Soc. Psychol. *70*, 913. https://doi.org/10.1037/0022-3514.70.5.913.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., and Baron, J. (2012). The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions (2012 AAAI Fall Symposium Series).

Wolf, M., Krause, J., Carney, P.A., Bogart, A., and Kurvers, R.H.J.M. (2015). Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PLoS One *10*, e0134269. https://doi.org/10.1371/journal.pone.0134269.

Wolf, M., Kurvers, R.H.J.M., Ward, A.J.W., Krause, S., and Krause, J. (2013). Accurate decisions in an uncertain world: collective cognition increases true positives while decreasing false positives. Proc. R. Soc. Lond. B *280*, 20122777. https://doi.org/10.1098/rspb.2012.2777.

Wood, G., Tyler, T.R., and Papachristos, A.V. (2020). Procedural justice training reduces police use of force and complaints against officers. Proc. Natl. Acad. Sci. U S A *117*, 9815–9821. https://doi.org/10.1073/pnas.1920671117.

Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., and Malone, T.W. (2010). Evidence for a collective intelligence factor in the performance of human groups. Science *330*, 686–688. https://doi.org/10.1126/science.1193147.

Yu, S., Pleskac, T.J., and Zeigenfuse, M.D. (2015). Dynamics of postdecisional processing of confidence. J. Exp. Psychol. Gen. *144*, 489. https://doi.org/10.1037/xge0000062.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R (version 4.0.4)	R-Core-Team, 2021	https://www.R-project.org/
Other		
Skin cancer dataset	Argenziano et al. (2003)	https://pnas.org/content/113/31/8777/tab-figures-data
Fingerprint dataset	Tangen et al. (2020)	https://osf.io/hgx3s
Forecasting dataset good Judgment project	Ungar et al. (2012)	https://dataverse.harvard.edu/dataverse/gjp
General knowledge dataset	Yu et al. (2015)	https://osf.io/cuzqm/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Ralf Kurvers (kurvers@mpib-berlin.mpg.de).

Materials availability

No materials were newly generated for this paper.

Data and code availability

- Data: The skin cancer dataset can be accessed at pnas.org/content/113/31/8777/tab-figures-data. The fingerprint dataset can be accessed at osf.io/hgx3s. The geopolitical dataset is part of the Good Judgment Project and accessible at dataverse.harvard.edu/dataverse/gjp. The general knowledge dataset is accessible at osf.io/cuzqm/. The Breast Cancer Surveillance Consortium (BCSC) holds legal ownership of the breast cancer dataset. Information regarding data requests can be found at bcsc-research.org/.
- Code: The code for reproducing the results and figures of the statistical argument (Figure 1) and the numerical simulations (Figure 2) are uploaded at the Open Science Framework: https://osf.io/rvdxz/. The code for reproducing the results and figures of the skin cancer dataset (Figures 3 and 6) are also uploaded here to illustrate how the results were calculated for one of the datasets. Simulations were done in R (version 4.0.4) (R-Core-Team, 2021).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This research did not conduct new experiments but analyzed existing datasets. Descriptions of the different datasets are provided below.

METHOD DETAILS

Breast cancer dataset

Full information on the breast cancer dataset can be found in Carney et al. (2012); we have also summarized the dataset in our previous publications (Kurvers et al., 2016, 2019; Wolf et al., 2015). The brief summary provided here largely adopts these earlier descriptions. Mammograms were randomly selected from screening examinations performed on women aged 40–69 years between 2000 and 2003 from US mammography registries affiliated with the Breast Cancer Surveillance Consortium (BCSC). 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 invited radiologists, 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 cancer patients (*n* = 27). Women who remained cancer-free for a period of two years were classified as noncancerous patients (*n* = 128; 17% prevalence).

Radiologists viewed the digitized images on a computer. They saw two images at the same time (left and right breasts) and were able to alternate quickly (≤1 s) between paired images, to magnify a selected part of an image, and to identify abnormalities by clicking on the screen. For each case, the craniocaudal and mediolateral oblique views of both breasts were presented simultaneously, followed by each view in combination with its prior comparison image. 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). The cases were evaluated 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 the same 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 in order to ensure that each radiologist evaluated the same set of cases, resulting in 15,655 diagnoses by 101 radiologists based on 155 cases. Between the two stages, radiologists were randomly assigned to one of three intervention treatments. Because there were no strong treatment differences (Geller et al., 2014), 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 (SCC) in Seattle and analyzed at the Max Planck Institute for Human Development (MPIB), Germany. Each registry, the SCC, and the MPIB received institutional review board approval for active and passive consent processes or were 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

Full information on the skin cancer dataset can be found in Argenziano et al. (2003); we have also summarized the dataset in our previous publications (Kurvers et al., 2015, 2016, 2019). The brief summary provided here largely adopts these earlier descriptions. The dataset comprises 4,320 diagnoses by 40 dermatologists of 108 skin lesions and was collected during a web-based consensus meeting. 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. Skin lesions were obtained from the Department of Dermatology, University Frederico 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 Skin and Cancer Associates (Plantation, Florida). The study was designed 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 dermatologists who participated in the study had a minimum of five years of experience in dermoscopy practice, teaching, and research. Dermatologists first received 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 use the diagnostic algorithm with the highest diagnostic accuracy, which is also the one most widely used for melanoma detection: pattern analysis. We treated the decision to classify a lesion as a melanoma as a positive test result.

Fingerprint analyses

Full information on the fingerprint analysis dataset can be found in Tangen et al. (2020); here, we provide a brief summary only. The full dataset comprises 1,728 evaluations by 36 professional fingerprint examiners of 48 fingerprint pairs. The fingerprint examiners were recruited from the Australian Federal Police, Queensland Police Service, Victoria Police, and New South Wales Police (mean experience = 16.4 years). Novices also participated in testing, but we used only data from the professional examiners in our analyses. The study was cleared by the ethical board of The University of Queensland and The University of Adelaide. Each of the 36 fingerprint examiners was presented with the same set of 24 fingerprint pairs from the same finger (targets) and 24 highly similar pairs from different fingers (distractors) in a different random order. Each pair consisted of a crime-scene "latent" fingerprint and a fully-rolled "arrest" fingerprint, and participants were asked to provide a rating on a 12-point scale ranging from 1 (Sure Different) to 12 (Sure Same). The distractors were created by running each latent fingerprint through the National Australian Fingerprint Identification System—which contains roughly 67 million fingerprints—to return the most similar exemplars from the database. On the first 44 of 48 trials (22 targets, 22 distractors), participants were given 20 s to examine the prints. On the final four trials (2 targets, 2 distractors), they had an unlimited amount of time to make a decision. For consistency, we excluded these last four trials, resulting in a total of 1,584 evaluations by 36 professional fingerprint examiners of 44 fingerprint pairs. To investigate the case of binary decision-making, we converted the answers from the 12-point scale into 'different' (6 and below) and 'same' (7 and above).

Forecasting dataset

The forecasting dataset is part of the Good Judgment Project (Ungar et al., 2012); we have previously summarized the dataset in (Kurvers et al., 2019). The brief summary provided here largely adopts this earlier description. The Good Judgment Project is a large-scale forecasting project, running over several years and with a wide variety of participants and settings (e.g., training schedules, team competitions). The subject pool consists of a mix of laypeople and geopolitical experts; participants were free to enter the forecasting competition. We used data from the first year of the forecasting project, when participants made 102 forecasts by answering 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 9,500 on 30 September 2011?" Participants were asked to estimate the probability of the future event on a scale from 0 to 1. We only included data from the individual condition (i.e. we excluded individuals who observed crowd information or participated in prediction markets). We excluded questions with more than two possible answers and questions for which the correct answer could not be irrefutably determined ($n = 8$), resulting in 94 questions. Finally, we excluded forecasters who answered fewer than 90 questions, resulting in 89 forecasters. The total dataset we used thus contained 8,258 forecasts by 89 forecasters on 94 geopolitical events. In some cases, forecasters updated their forecasts over time, thereby giving multiple responses. In such cases, we used their first forecast only. To investigate the case of binary decision-making, we converted the probability scores into two categories: 0 (probabilities <0.5) and 1 (probabilities >0.5). Scores of 0.5 were randomly converted to either 0 or 1. Figure S6 presents the results of additional analyses using the probability forecasts directly without any conversion, with the Brier score as the accuracy measure.

General knowledge dataset

This dataset is based on Study 3 in Yu et al. (2015); we have previously summarized the dataset in Kurvers et al. (2019). The brief summary provided here largely adopts the earlier description. The dataset contains binary responses to the question, "Which of the following two cities has more inhabitants?" The stimulus set consisted of 1,000 randomly generated pairs of cities from a list of the 100 most populous cities in the United States in 2010 as determined by the US Census Bureau. After seeing a fixation cross, participants observed a pair of cities and after 1.6 s, they were cued to decide. Participants rated 1,000 pairs in two sessions. Participants ($n = 109$) were recruited from the Michigan State University (MSU) psychology research participant pool and received class credits plus a \$0–\$4 bonus per session. Participants gave informed consent according to the MSU Institutional Review Board guidelines. We excluded participants who did not complete both sessions, resulting in 99 participants, all of whom provided a decision on each of the 1,000 city pairs.

QUANTIFICATION AND STATISTICAL ANALYSIS

Numerical simulations: Correlated errors

To study the effect of correlations between the decisions of different individuals in the numerical simulations, we used an "opinion leader" approach (Grofman et al., 1983; Kurvers et al., 2019), fixing the number of cases m to 25. One of the n individuals of a group was assigned to be the ''opinion leader''. The opinion leader made independent decisions for the m cases based on its (sampled) accuracy level. The decisions of all the remaining individuals were made dependent on the opinion leader's sequence, according to a correlation parameter p_c ($0 \leq p_c \leq 1$). This correlation parameter governs the level of correlation between decision makers, ranging from 0 (maximum amount of independence) to 1 (maximum amount of dependence). Each of the remaining individuals first made m independent decisions based on its (sampled) accuracy level. Next, the sequence of decisions was made dependent on the opinion leader's sequence using the correlation parameter p_c. For each case, starting at case i = 1, we randomly selected with probability $(1 - p_c)$ a decision from the set of remaining decisions from that individual (i.e. decisions from cases j that have not yet been selected, $j \geq i$), and we took with probability p_c the same decision as the decision of the opinion leader from this set. If the same decision was not present in that individual's set of remaining decisions, we randomly selected a decision from this set. We then moved on to the next case i + 1. We repeated the numerical simulations at three levels of p_c: 0 (complete independence, the assumption used in Figure 2), 0.5, and 1 (maximum level of correlation). Note that even if p_c = 1, there can still be disagreement between individuals in a group, namely, when the numbers of 0s and 1s in their respective vectors are not equal. The results of this analysis are shown in Figure S4.

Calculation of discrimination ability and criterion

For the simulations of the empirical data, we repeatedly and randomly drew groups of different sizes. The decisions of the individuals drawn were combined using the majority rule. Based on individual or group decisions, we calculated the number of hits (H), misses (M), false positives (FP), and correct rejections (CR). From this we calculated the hit rate (HR) as HR = H/(H + M) and the false-positive rate (FPR) as FPR = FP/(FP + CR). From this we calculated discrimination ability d' and criterion c using:

$$d' = z(HR) - z(FPR) \text{ and}$$

$$c = -(z(HR) + z(FPR))/2.$$

where $z(.)$ is the inverse distribution function of the normal distribution, yielding the z-scores for the hit rate (HR; also known as sensitivity) and the false-positive rate (FPR; 1 – specificity). We followed the standard approach of adding 0.5 to each of the four categories' counts (i.e. H, M, FP, and CR) to avoid the possibility of infinite values (Macmillan and Creelman, 2005).