Flexible human collective wisdom

Mordechai Z. Juni[1] & Miguel P. Eckstein[1,2]

[1]Department of Psychological & Brain Sciences, University of California, Santa Barbara

[2]Institute for Collaborative Biotechnologies, University of California, Santa Barbara

Citation:

Juni, M. Z., & Eckstein, M. P. (2015). Flexible human collective wisdom. *Journal of Experimental Psychology: Human Perception and Performance.* Advance online publication. http://dx.doi.org/10.1037/xhp0000101

Abstract

Group decisions typically outperform individual decisions. But how do groups combine their individual decisions to reach their collective decisions? Previous studies conceptualize collective decision-making using static combination rules, be it a majority-voting rule or a weighted averaging rule. Unknown is whether groups adapt their combination rules to changing information environments. We implemented a novel paradigm for which information obeyed a mixture of distributions, such that the optimal Bayesian rule is non-linear and often follows minority opinions, while the majority rule leads to suboptimal but above chance performance. Using perceptual (Exp1) and cognitive (Exp2) signal detection tasks, we switched the information environment halfway through the experiments to a mixture of distributions without informing participants. Groups gradually abandoned the majority rule to follow any minority opinion advocating signal presence with high confidence. Furthermore, groups with greater ability to abandon the majority rule achieved higher collective-decision accuracies. Importantly, this abandonment was *not* triggered by performance loss for the majority rule relative to the first half of the experiment. Our results propose a new theory of human collective decision-making: Humans make inferences about how information is distributed across individuals and time, and dynamically alter their joint decision algorithms to enhance the benefits of collective wisdom.

Keywords: group decision making; signal detection theory; ideal observer analysis; Bayesian optimal; wisdom of crowds

Self-organizing social animals employ democratic-like group decisions to choose foraging locations (Beckers, Deneubourg, & Goss, 1992; Prins, 1996; Seeley, Camazine, & Sneyd, 1991), navigation directions (Simons, 2004; Ward, Sumpter, Couzin, Hart, & Krause, 2008), nesting sites (Pratt, Mallon, Sumpter, & Franks, 2002; Seeley & Buhrman, 1999), camping sites (cf. Lewis & Clark expedition (Hastie & Kameda, 2005)), whether to relocate (Stewart & Harcourt, 1994; Sueur & Petit, 2008), and whether to wage war (Boehm, 1996). The principle of "one individual, one vote" is routinely observed throughout the animal kingdom (Conradt & Roper, 2003, 2005; Couzin, Krause, Franks, & Levin, 2005). Human groups regularly employ the majority rule (Kalven & Zeisel, 1966; Kameda, Tindale, & Davis, 2003; Tindale, 1989), perhaps because it is quite effective (cf. Condorcet's jury theorem (Condorcet, 1785)) while taking little cognitive and social effort to implement (Hastie & Kameda, 2005). Moreover, the majority rule can often approach the performance accuracy of optimal (but computationally difficult) combination rules (Eckstein et al., 2012; Sorkin, Luan, & Itzkowitz, 2008). In special circumstances, such as two-member groups, people employ other combination rules such as weighted averaging (Bahrami et al., 2010). But when the majority rule is available, people tend to rely on it (Denkiewicz, Rączaszek-Leonardi, Migdał, & Plewczynski, 2013).

Previous studies have evaluated scenarios for which utilizing the majority-voting rule does not lead to great cost in performance accuracy relative to the optimal rule (Sorkin, West, & Robinson, 1998). Yet the same aggregation algorithm could be very effective in some information environments but not others (Koriat, 2012). Furthermore, previous studies have evaluated scenarios in which the distribution of information (stimulus strength) across observers is temporally invariant, and thus groups have no need to adapt their aggregation algorithm over the course of the experiment. Humans working alone adaptively select from a repertoire of decision strategies to improve their personal performance in different information environments (Rieskamp & Otto, 2006). The question remains whether humans working together adapt their group decision rules (cf. social decision schemes (Davis, 1973)) to improve the accuracy of their collective decisions in different information environments.

To explore this question of *group adaptability*, we implemented mixture of distributions scenarios where the quality of information (stimulus strength) was uneven across group-members and time, with different group-members receiving high quality information on different trials. Thus, on each signal-present trial, a different random observer received high quality

information while the other observers received low quality information. In such information environments, the optimal Bayesian rule (Geisler, 2003; Green & Swets, 1966; Knill & Richards, 1996; Peterson, Birdsall, & Fox, 1954) with knowledge of how information might be distributed across individuals and time is non-linear and often follows minority opinions, while the majority rule leads to suboptimal but above chance performance. We evaluate (i) whether human groups stop using the majority rule in the mixture of distributions scenarios even though they are not informed of the change in the information environment; (ii) various group decision rules to assess what integration algorithm human groups might employ instead; and (iii) whether groups that are better able to alter their collective integration rules achieve higher group-decision accuracies.

In addition, previous studies comparing human collective integration rules to optimal (Sorkin, Hays, & West, 2001) have evaluated simpler information environments where the optimal Bayesian integration model reduces to a weighted linear model (Sorkin & Dai, 1994). Here, we evaluate human collective integration rules for more complex information environments (mixture of distributions) where the optimal Bayesian integration model is non-linear. We assess (i) how human group decisions differ from the optimal Bayesian non-linear rule on a trial-by-trial basis; and (ii) try to account for the human departures from the optimal Bayesian model in terms of a human expectation of future possible changes in the information environment.

The experiments consisted of perceptual (Exp1) and cognitive (Exp2) yes/no signal detection tasks (Green & Swets, 1966) (50% signal trials, 50% noise trials) with similar structure in the statistical distribution of information. During the first half of the experiments, the stimulus strength (characterized by the signal to noise ratio, which is the distance in standard deviation units between two equal variance Gaussian distributions) was the same across all group members (*equal strength condition*). In such circumstances, the majority-voting rule incurs only modest accuracy costs relative to the optimal integration rule. Halfway through the experiments, and unknown to participants, the statistical distribution of information changed so that on each signal trial a different random member of the group received strong stimulus-strength while the other two members received weak stimulus-strength (*mixture condition*). In such circumstances, the majority-voting rule incurs substantial accuracy costs (but above chance performance) relative to the optimal integration rule.

The perceptual task (Exp1) was to detect on each trial whether or not there was a spatial Gaussian-shaped luminance signal embedded in additive luminance white noise (same noise sample across observers). The cognitive task (Exp2) was to predict on each trial whether or not a sandstorm would occur based on fictitious environmental measurements sampled from univariate Gaussian distributions (partially correlated across observers[1]). The experimental sessions had 200 equal strength condition trials followed by 200 mixture condition trials. Figure 1 shows a schematic overview of the experimental tasks. On each trial, participants first recorded their personal confidence using a 10-point rating scale, then took turns announcing to the group their personal confidence rating, and then conferred with each other until they reached and agreed what their group response should be, which was followed by feedback. Figure 2 shows a schematic of the response sequence.

[1] Correlation of random variables across observers = .2. This correlation was introduced to approximately match the average correlation between observer ratings in the perceptual task and thus allow meaningful comparisons across perceptual and cognitive tasks.

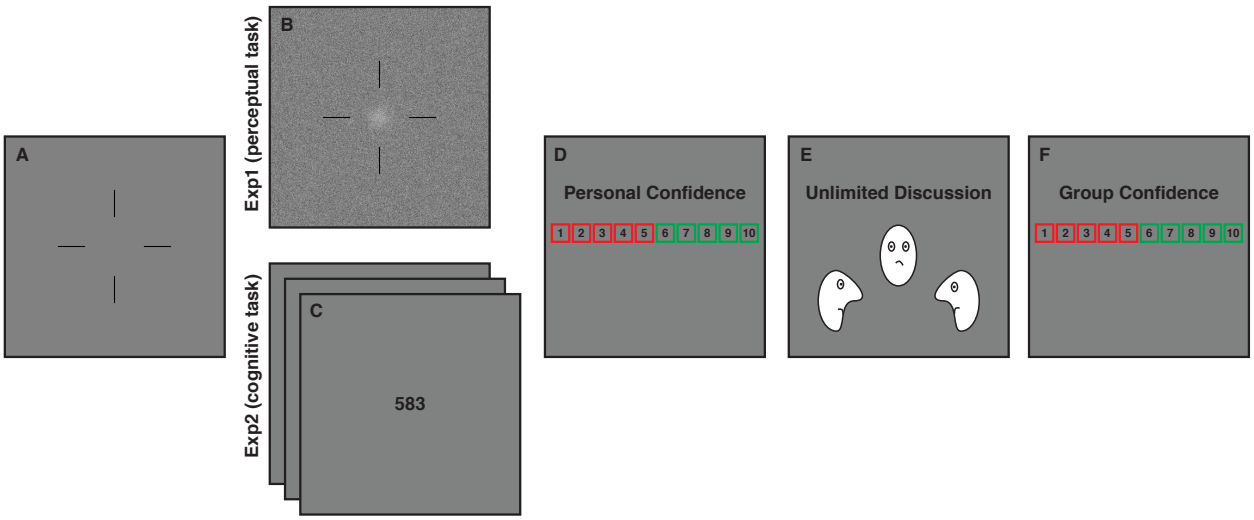

Figure 1. Schematic overview of the yes/no signal detection tasks (50% signal trials, 50% noise trials). A. Fixation cross at the start of the trial to direct attention to the center of the screen. **B.** In the perceptual task (Exp1), all group-members saw a grainy image with identical additive white noise. Signal strength was manipulated by changing the contrast of the Gaussian-shaped luminance blob that could be embedded at the center of the image. The contrast of the signal (when present) was either identical in all three images (equal strength condition), or stronger in the image of one random observer and weaker in the images of the two other observers (mixture condition). The identity of the observer receiving the strong signal was randomized and unknown throughout the mixture condition. **C.** In the cognitive task (Exp2), each group member saw a different set of Gaussian random variables (fictitious environmental measurements) that are predictive of sandstorms. Signal strength was manipulated by shifting the underlying distribution for signal-trial measurements closer to and further away from the underlying distribution for noise-trial measurements, thus inducing weaker and stronger signal-strengths respectively. The distance between the signal distribution and the noise distribution was either identical for all three observers (equal strength condition), or further apart for one random observer and closer together for the two other observers (mixture condition). The identity of the observer receiving the strong signal was randomized and unknown throughout the mixture condition. **D-F.** On each trial, participants first took turns announcing to the group their personal confidence rating, and then conferred with each other until they reached and agreed what their group response should be, which was followed by feedback.

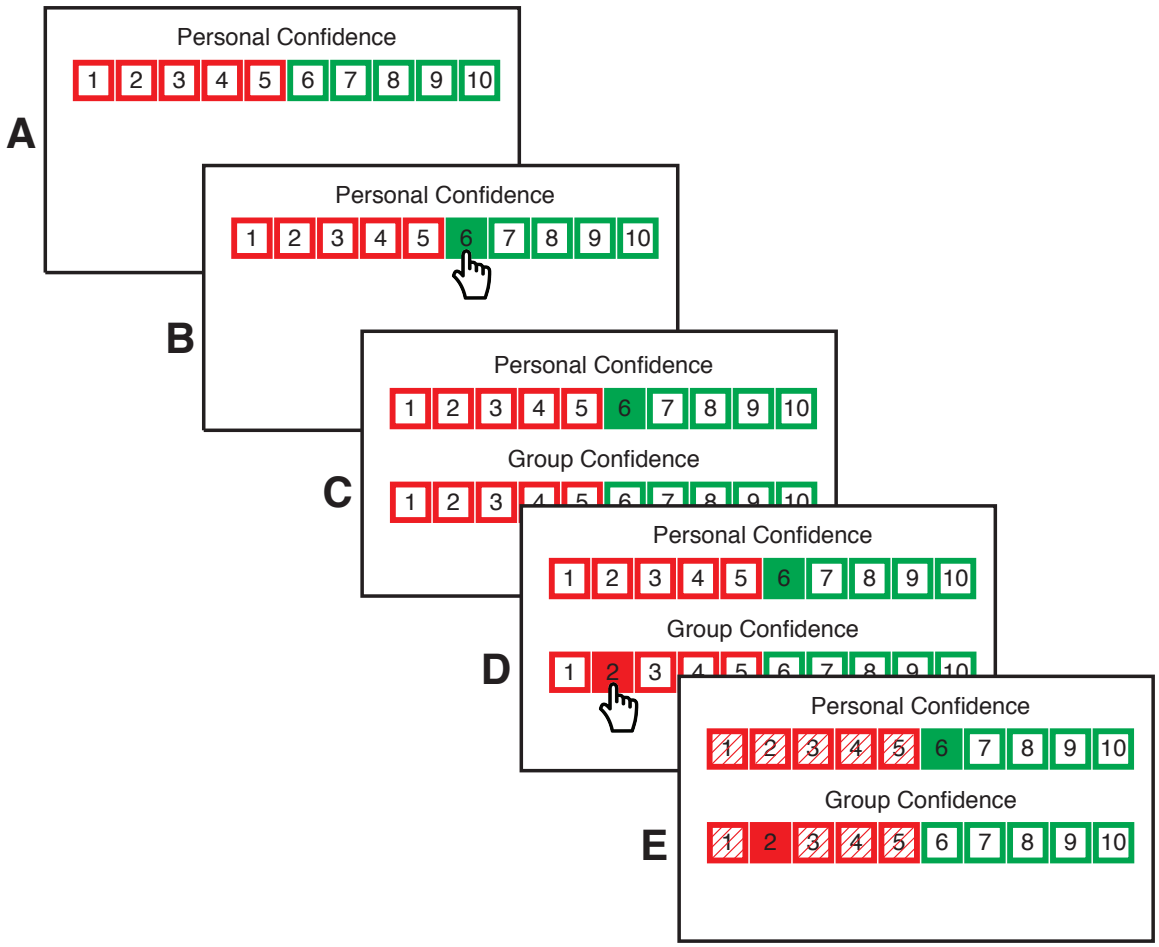

Figure 2. Schematic of the response sequence during the experimental phase.
A. Participants recorded their response using this colorful 10-point confidence scale. To respond "no", participants clicked on one of the five ratings in the red squares ranging from *most* confident 1 to *least* confident 5. To respond "yes", participants clicked on one of the five ratings in the green squares ranging from *least* confident 6 to *most* confident 10. **B**. In this example, the participant responded "yes" by clicking on "6", and the square was filled in to mark the response. **C.** When the participant finalized his or her response (they were allowed to change their response until they locked it in by pressing space bar), the group confidence scale appeared right below and the participants took turns announcing to the group their personal confidence rating for that trial (see Group Decision section under General Methods for details on how we maintained independence between participants). **D.** After discussing and agreeing on a group response for that trial, each participant clicked accordingly to record said response. In this example the group chose to respond "no" by clicking on "2", and the square was filled in to mark the group response. **E.** After pressing space bar to receive feedback, the ratings that count as a correct response for that trial were lightly shaded. In this example, the red squares were lightly shaded because it was a noise trial and the correct response was to say "no". This participant was wrong in his or her personal response, but the group's collective response was correct. To start the next trial, each participant pressed space bar whenever he or she was ready.

Experimental Methods

General Methods

Participants. Sixty undergraduate students at the University of California, Santa Barbara volunteered or earned partial course credit for their participation. All were naïve to the purpose of the study, and each participated in only one experiment. Each experiment had 10 groups. Each group consisted of three students participating together in the same room. Each participant worked on a separate computer, and dividers were in place so that one could not see the other two computers. The experiments were run using MATLAB and the Psychophysics Toolbox libraries (Brainard, 1997; Pelli, 1997).

Yes/No Task. In Experiment 1 (perceptual task) participants were informed that the Gaussian luminance signal would be present 50% of the time. In Experiment 2 (cognitive task) participants were informed that there would be a sandstorm 50% of the time. In both experiments participants responded "no" (i.e., *signal absent* or *no sandstorm*) and "yes" (i.e., *signal present* or *yes sandstorm*) using a colorful 10-point confidence scale, where 1 indicates a very high confidence that it was a noise trial and 10 indicates a very high confidence that it was a signal trial.

Group Decision. On each trial of the experimental phase, participants first recorded their personal confidence *before* announcing it to the group and talking with each other to choose the group confidence. To maintain independence between participants and ensure that everyone expressed aloud their true personal confidence, participants were instructed to (i) refrain from talking until *all three participants* announced that they were “done” recording their personal confidence and (ii) take turns announcing their personal confidence *before* discussing what their group confidence should be for that trial. The experimenter remained in the room to ensure that participants followed these two rules.

Methods for Experiment 1 (perceptual task)

Task. Participants were challenged to detect, both individually and collectively, whether or not there was a Gaussian luminance signal present in a grainy image with additive white noise (signal present 50% of the time). The experiment was carried out in a dimly lit room with each

participant viewing one of three different CRT monitors that were linearly calibrated to a mean luminance of approximately 28 cd/m^2. The resolution of each monitor was set to 800 by 600 pixels, and each pixel subtended approximately 0.037°. The familiarization and training phases took place in a single session that lasted between 1 and 1.5 hours. The experimental phase took place in a second session that lasted between 2.5 and 3 hours and was carried out between 1 and 3 days later (except for one group that carried out the second session 9 days later because one of the participants got ill in the interim).

Stimulus. Stimuli were 8-bit grey scale images that were shown at the center of the screen. The noise in each stimulus consisted of a 15° by 15° patch of Gaussian white noise (μ = 28 cd/m^2; σ = 4.375 cd/m^2; noise root-mean-square contrast = σ/μ = .1562). The signal (when present) consisted of a Gaussian blob (standard deviation = 0.5°) that was added to the center of the noise patch. The energy of the signal, denoted E, is defined as the sum of the squared luminance values of the entire Gaussian blob as follows:

$$E = \sum_{x}^{X} \sum_{y}^{Y} S(x,y)^2 \ , \qquad (1)$$

where S(x,y) is the luminance value of the Gaussian blob at each pixel location. The signal to noise ratio (SNR) in the perceptual task is the distance in standard deviation units between an ideal observer's respective decision-variable distributions for signal-present images and signal-absent images. For white noise, it can be calculated from the signal and noise as follows (Burgess, Wagner, Jennings, & Barlow, 1981; Watson, Barlow, & Robson, 1983):

$$\text{SNR} = \frac{\text{root signal energy}}{\text{noise standard deviation}} = \frac{\sqrt{E}}{\sigma} \ . \qquad (2)$$

The standard deviation of the white noise was the same throughout the experiment. The energy of the signal was manipulated by changing the contrast of the Gaussian blob.

Familiarization phase (850 trials). Each participant completed this alone as there was no group decision component during this phase. The stimulus was displayed for 500 msec. Participants indicated with a key press whether they thought that the signal was absent or present, and they heard a beep (through headphones) when they responded incorrectly. They were also informed that as the familiarization phase progressed, it would get harder and harder to detect the signal

and that they would start hearing the beep more and more often. During the first block (50 trials), the peak signal contrast (when present) was 50%. During the second block (50 trials), it was lowered to 25%, and during the third block (50 trials) it was lowered to 15%. From the fourth block onwards, it was lowered every two blocks (100 trials) from 10% to 8.5% to 7% to 5.5% to 4% to 2.5% and finally to 1%.

Training phase (200 trials). Each participant completed this alone as there was no group decision component during this phase. The stimulus was displayed for 500 msec, and the peak signal contrast (when present) was 2% throughout this phase. Participants responded using the 10-point confidence scale, and they received feedback at the end of each trial.

Experimental phase (400 trials). Each experimental trial had an individual decision component and a group decision component. The stimulus was displayed for 500 msec. Each stimulus contained a unique random noise patch that was identical across all three computer displays. During the first 200 trials of the experimental phase (equal strength condition), the peak signal contrast (when present) was 2% (SNR = 3.07) on all three computer displays. During the last 200 trials of the experimental phase (mixture condition), the peak signal contrast (when present) was 0.5% (SNR = 0.77) on two of the computer displays (randomized on each signal trial) and 9% (SNR = 13.8) on the remaining computer display. Keep in mind that all three participants saw identical noise patches during these signal-present trials; the only thing that was different across computer displays was the contrast strength of the signal. Participants responded using the 10-point confidence scale, and they received feedback at the end of each trial. The stimuli and presentation order were kept the same across all ten groups.

Methods for Experiment 2 (cognitive task)

Task. Participants were challenged to predict, both individually and collectively, whether or not there would be a sandstorm on each trial (sandstorm present 50% of the time). The experiment was carried out in a well-lit room with each participant viewing one of three different LCD monitors. The entire experiment was carried out in a single session that lasted between 2.5 and 3 hours.

Participants were told that each computer would provide fictitious measurements of some unknown environmental factor that is predictive of sandstorms (e.g., wind-speed, atmospheric pressure, humidity, and temperature). Participants were informed that each computer was a device that measured a different environmental factor and that it was futile to compare with each other the actual measurements that they received (this was done to maintain independence between participants). Instead, they were advised that, while they could communicate with each other however they wished (including sharing their measurements if they really wanted to), the easiest way to communicate would be to convert their measurements into a common confidence scale that would range between 1 (very confident "no" sandstorm) and 10 (very confident "yes" sandstorm).

To learn about their device (see familiarization phase below), participants were given an opportunity to observer color-coded measurements that were taken by their device both during sandstorms (in green) and in the absence of sandstorms (in red). They were informed that just like there is overlap in the distribution of heights between women and men, so too there is overlap in the distribution of green and red measurements because (1) the devices are imperfect and have some degree of random measuring error and (2) the environmental factors are not perfectly predictive of sandstorms. Their job, however, was to learn as best they could what numbers are associated with the presence of a sandstorm ("yes" sandstorm) and what numbers are associated with the absence of a sandstorm ("no" sandstorm). In addition, to facilitate learning, they were informed that, just like it is more likely to be a man the taller the height, so to it is more likely to be a sandstorm the higher the measurement.

During the actual trials, each computer displayed three measurements one after another, and participants had to predict whether or not there would be a sandstorm. Participants were told that they would each be provided with a triplet of measurements instead of one single measurement to compensate for the fact that the devices have some degree of random measuring error. The three measurements, or numbers, were colored in black, and participants were told that, essentially, they had to guess if the three numbers were drawn from the distribution of red numbers or from the distribution of green numbers (i.e., "Should the numbers have been colored in red or in green?"). When communicating with each other, participants would sometimes use the sandstorm analogy (e.g., "So do you really think it's a sandstorm?") and sometimes the color analogy (e.g., "I'm pretty sure it is red").

Stimulus. Each experimental stimulus consisted of three Gaussian random variables (rounded to their nearest integer) that were displayed for 250 msec each (with a 250 msec gap between numbers). The three random variables for each stimulus were drawn from one of four statistically independent Gaussian distributions. The four Gaussian distributions had the same standard deviation (σ_0) but different means depending on whether it was a noise trial (μ_{NOISE}), a weak-signal trial (μ_{WEAK}), a medium-signal trial (μ_{MEDIUM}), or a strong-signal trial (μ_{STRONG}). The parameters of the distributions were pseudo-randomly chosen so that the signal to noise ratio (SNR_0) of weak-signal, medium-signal, and strong-signal random variables would be 0.29, 0.78, and 3.46, respectively. The signal to noise ratios in the cognitive task are the respective distances (in standard deviation units) between the three signal distributions and the noise distribution. The signal to noise ratio is thus equivalent to d-prime, denoted d', which is a measure of sensitivity commonly used in signal detection theory and defined as follows (Green & Swets, 1966):

$$\mathrm{SNR}_0 = \mathrm{d}'_0 = \frac{\mu_{SIGNAL} - \mu_{NOISE}}{\sigma_0} \ . \tag{3}$$

As each stimulus consisted of a triplet of measurements (i.e., three random draws from the same distribution), the effective signal to noise ratio (SNR_{eff}) of the ideal observer takes into account the standard deviation of the mean of the triplet of measurements: $\sigma_{\mathrm{eff}} = \frac{\sigma_0}{\sqrt{n}}$. Hence, the ideal observer SNR_{eff} is given as follows:

$$\mathrm{SNR}_{\mathrm{eff}} = \mathrm{SNR}_0 \sqrt{n} \ , \tag{4}$$

where n = 3 in our case. This means that the SNR_{eff} of weak-signal triplets, medium-signal triplets, and strong-signal triplets was 0.5, 1.35, and 6, respectively.

The parameter values of the four Gaussian distributions for each computer were as follows (for clarity, we report the parameter values rounded to their nearest integer):

Comp1: $\sigma_0 = 55;\ \mu_{NOISE} = 286;\ \mu_{WEAK} = 301;\ \mu_{MEDIUM} = 328;\ \mu_{STRONG} = 476.$
Comp2: $\sigma_0 = 69;\ \mu_{NOISE} = 360;\ \mu_{WEAK} = 380;\ \mu_{MEDIUM} = 414;\ \mu_{STRONG} = 600.$
Comp3: $\sigma_0 = 89;\ \mu_{NOISE} = 460;\ \mu_{WEAK} = 486;\ \mu_{MEDIUM} = 529;\ \mu_{STRONG} = 767.$

The reason we used three different sets of parameter values (one set for each computer and thus participant) was to ensure independence between the participants in case they decided to share their measurements with each other. Hence, we emphasized to them that their respective

measurements were being drawn from different regions of the number line and that it would be futile to share their numbers with each other.

The values of the three μ_{NOISE} parameters (one for each computer and thus participant) were the only ones that were picked pseudo-randomly. Specifically, they were randomly drawn from the following bounded uniform distributions: Comp1: (260,290); Comp2: (350,380); Comp3: (440,470). All other parameter values were then determined automatically as follows: μ_{MEDIUM} was set to be 15% greater than μ_{NOISE}; σ_0 was set so that the SNR_{eff} of medium-signal triplets would be 1.35; μ_{WEAK} and μ_{STRONG} were set so that the respective SNR_{eff} of weak-signal triplets and strong-signal triplets would be 0.5 and 6. While each computer's random variables were drawn from different distributions, we ensured that every single number that was displayed had exactly three digits (the lowest number displayed on any of the three computers was 117 and the highest was 947).

Lastly, while the three random variables of each stimulus were uncorrelated (i.e., the first, second, and third measurements of each triplet were independently drawn), there was a correlation *across* the three computers. The first measurements of each trial (i.e., the first measurement on Comp1, the first measurement on Comp2, and the first measurement on Comp3) were correlated with each other; and so were the second measurements on each trial and so were the third measurements on each trial. During the training phase, when participants were only making decisions individually, the correlation between the random variables *across* the three computers was 1 (i.e., they were perfectly correlated). During the experimental phase, when participants were also making group decisions, the correlation between the random variables *across* the three computers was .2. This correlation was introduced to approximately match the average correlation between observer ratings in Experiment 1 and thus allow meaningful comparisons across perceptual and cognitive tasks.

Familiarization phase. Each participant observed 300 color-coded numbers and completed this alone as there was no group decision component during this phase. Each number was displayed at the center of the screen for 250 msec, with a 250 msec gap between numbers. Hence this familiarization phase was very quick and lasted 2.5 min. Each participant controlled for him or herself at all times whether they wanted to see green numbers (i.e., numbers that were measured

during a sandstorm) or red numbers (i.e., numbers that were measured in the absence of a sandstorm). Green numbers were drawn from a Gaussian with μ_{MEDIUM}, while red numbers were drawn from a Gaussian with μ_{NOISE}. The numbers were shown one after another very rapidly, and participants switched between green numbers and red numbers by pressing a key ("Y" for green numbers; "N" for red numbers). They were advised to take their time and to not switch back and forth between the colors too rapidly so that they could pay attention and learn the respective distributions as well as possible.

Training phase (200 trials). Each participant completed this alone as there was no group decision component during this phase. The stimulus on each trial consisted of three numbers that were colored in black. The three numbers were displayed one after another at the center of the screen for 250 msec, with a 250 msec gap between numbers. The three numbers on each trial were drawn either from a Gaussian with μ_{NOISE} or from a Gaussian with μ_{MEDIUM} depending on whether it was a noise trial (i.e., no sandstorm) or a signal trial (i.e., yes sandstorm). Participants responded using the 10-point confidence scale, and they received feedback at the end of each trial.

Experimental phase (400 trials). Each experimental trial had an individual decision component and a group decision component. The stimulus on each trial consisted of three numbers that were colored in black. The three numbers were displayed one after another at the center of the screen for 250 msec, with a 250 msec gap between numbers. During the first 200 trials of the experimental phase (equal strength condition), the three numbers on each trial were drawn either from a Gaussian with μ_{NOISE} or from a Gaussian with μ_{MEDIUM} depending on whether it was a noise trial (i.e., no sandstorm) or a signal trial (i.e., yes sandstorm). During the last 200 trials of the experimental phase (mixture condition), the three numbers were drawn from a Gaussian with μ_{NOISE} on all three computers during noise trials, as opposed to signal trials when they were drawn from a Gaussian with μ_{WEAK} on two computers (randomized on each signal trial) and a Gaussian with μ_{STRONG} on the remaining computer.

The numbers on each computer were kept the same across all ten groups, and the trial order for each computer was identical to its corresponding order during the experimental phase of Experiment 1 as follows: μ_{NOISE} corresponds to signal absent; μ_{WEAK} corresponds to 0.5% contrast signal; μ_{MEDIUM} corresponds to 2% contrast signal; and μ_{STRONG} corresponds to 9% contrast signal. Participants responded using the 10-point confidence scale, and they received feedback at the end of each trial.

Proportion Correct and Choice Probability

Participants were instructed to respond "no" by clicking on one of the red-colored boxes numbered 1 through 5, and to respond "yes" by clicking on one of the green-colored boxes numbered 6 through 10 (see Figure 2). In our analyses we looked primarily at two different measures: Proportion Correct and Choice Probability.

Proportion Correct. We measured this for individuals, groups, and various group decision rules. For individuals, *proportion correct* is the proportion of trials that the participant's personal yes/no response is correct. For groups, *proportion correct* is the proportion of trials that the group's collective yes/no response is correct. For the various group decision rules, *proportion correct* is the proportion of trials that the rule's yes/no response is correct.

Choice Probability. We measured this for each group decision rule. The *choice probability* of each rule is the proportion of trials that the rule's yes/no response is the same as the group's collective yes/no response (irrespective of whether or not that yes/no response is correct).

Group Decision Rules (brief descriptions)

We evaluated eight different collective integration algorithms. Each rule makes a yes/no response on each trial by combining the group-members' personal confidence ratings on that trial. See Table 1 for brief mathematical expressions, and see Appendix A for full descriptions and mathematical expressions.

1. The *majority* rule responds "no" if a majority of the group-members' personal confidence ratings on that trial are between 1 and 5. Conversely, this rule responds "yes" if a majority of the group-members' personal confidence ratings on that trial are between 6 and 10. Note that the majority rule never follows a minority opinion no matter how confident that opinion may be, and that it never goes against a unanimous opinion.

2. The *majority with exceptions* rule follows the majority rule, except on trials when one (or more) of the group-members' personal confidence ratings is a highly confident "yes", in which case this rule responds "yes" even if the other group-members' personal confidence ratings on that trial are between 1 and 5. Note that while this rule follows a minority opinion endorsing signal-presence with high confidence, this rule *does not* follow a minority opinion endorsing *signal-absence* with high confidence. In addition, note that just like the majority rule, this rule never goes against a unanimous opinion.

3. The *averaging* rule responds "no" if the average of the group-members' personal confidence ratings on that trial is below criterion (see Criterion section in Appendix A for details). Conversely, this rule responds "yes" if the average of the group-members' personal confidence ratings on that trial is above criterion.

4. The *weighted linear combination* rule is similar to the averaging rule, except that it differentially weights participants' personal confidence ratings based on the covariance of group-members' personal confidence ratings across all trials, and how well each group-member discriminates between signal and noise trials. If the weighted average is below criterion it responds "no", and if the weighted average is above criterion it responds "yes".

5 and 6. The *optimal Bayesian* model (*linear* for the equal strength condition; *non-linear* for the mixture condition) is afforded knowledge about all possible signal-strengths and the statistical distribution of information across group-members and time, which it uses to compute on each trial the likelihood of jointly eliciting the three *personal* ratings of that trial given that it is a signal (or noise) trial. If the likelihood is below criterion it responds "no", and if the likelihood is above criterion it responds "yes". The optimal linear model for the equal strength condition is essentially the same as the weighted linear combination rule. The optimal non-linear model for the mixture condition considers the three actual combinations of signal-strength assignments to group members that are possible during signal trials: (i) weak, weak, strong; (ii) weak, strong, weak; and (iii) strong, weak, weak.

7. The *Bayesian-with-uncertainty* model (for the mixture condition) is not afforded full knowledge about the signal-strengths that are possible when it computes the likelihood of jointly eliciting the three *personal* ratings of the current trial given that it is a signal (or noise) trial. Hence, it considers many different combinations of signal-strength assignments to group members, instead of the three actual sets that the optimal non-linear model considers. This non-linear Bayesian model with signal-strength uncertainty considers 15^3 (i.e., 3375) different combinations of signal-strength assignments to group members.

8. The *associative heuristic* rule (for the mixture condition) responds "no" except on trials when one (or more) of the group-members' personal confidence ratings is a highly confident "yes", in which case this rule responds "yes" no matter what the other group-members' personal confidence ratings are on that trial. Similar to the majority with exceptions rule, this rule follows a minority opinion endorsing signal-presence *with high confidence*. But importantly, and unlike the majority with exceptions rule, this rule responds "no" and goes against a majority or unanimous opinion endorsing signal-presence *with low confidence* (i.e., this rule responds "no" when no-one's personal confidence rating is a highly-confident "yes").

Table 1. Group Decision Rules (brief mathematical expressions)

Descriptive title	**Collective integration algorithm (10-point rating scale)**	**Group response** 0 = "signal absent" 1 = "signal present"
Majority	$\hat{x} = \sum_{j}^{n} step\left(x_j\right)$; where $step(x) = \left\{ \begin{array}{l} +1, \text{ if } x > 5.5 \\ -1, \text{ if } x < 5.5 \end{array} \right\}$	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < 0 \\ 1, \text{ if } \hat{x} > 0 \end{array} \right\}$
Majority with exceptions	$\hat{x} = \sum_{j}^{n} step\left(x_j\right)$; where $step(x) = \left\{ \begin{array}{l} \infty, \text{ if } x \geq k \\ +1, \text{ if } 5.5 < x < k \\ -1, \text{ if } x < 5.5 \end{array} \right\}$ k = 7, 8, 9, or 10, and is fit separately for each group	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < 0 \\ 1, \text{ if } \hat{x} > 0 \end{array} \right\}$
Averaging	$\hat{x} = \frac{1}{n} \sum_{j}^{n} x_j$	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < criterion \\ 1, \text{ if } \hat{x} > criterion \end{array} \right\}$ *criterion* is fit separately for each group
Weighted Linear	$\hat{x} = \sum_{j}^{n} w_j x_j$; where $[w_1, w_2 \ldots w_n]^{\mathrm{T}} = \boldsymbol{\Sigma}^{-1}(\mu_{SIGNAL} - \mu_{NOISE})^{\mathrm{T}}$	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < criterion \\ 1, \text{ if } \hat{x} > criterion \end{array} \right\}$ *criterion* is fit separately for each group
Optimal Bayesian linear (equal strength condition)	$\hat{x} = \frac{\mathrm{P}\left[X \mid signal\right]}{\mathrm{P}\left[X \mid noise\right]} = \frac{\frac{1}{Q}\exp\left(-\frac{1}{2}(X - \mu_{SIGNAL})\sum^{-1}(X - \mu_{SIGNAL})^{\mathrm{T}}\right)}{\frac{1}{Q}\exp\left(-\frac{1}{2}(X - \mu_{NOISE})\sum^{-1}(X - \mu_{NOISE})^{\mathrm{T}}\right)}$	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < criterion \\ 1, \text{ if } \hat{x} > criterion \end{array} \right\}$ *criterion* is fit separately for each group
Optimal Bayesian non-linear (mixture condition)	$\hat{x} = \frac{\mathrm{P}\left[X \mid signal\right]}{\mathrm{P}\left[X \mid noise\right]} = \frac{\frac{1}{n}\sum_{j}^{n}\frac{1}{Q}\exp\left(-\frac{1}{2}(X - \mu_{\text{j-Strong}})\sum^{-1}(X - \mu_{\text{j-Strong}})^{\mathrm{T}}\right)}{\mathrm{P}\left[X \mid noise\right]}$	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < criterion \\ 1, \text{ if } \hat{x} > criterion \end{array} \right\}$ *criterion* is fit separately for each group
Bayesian with uncertainty (mixture condition)	$\hat{x} = \frac{\mathrm{P}\left[X \mid signal\right]}{\mathrm{P}\left[X \mid noise\right]} = \frac{\frac{1}{15^n}\sum_{h}^{15^n}\frac{1}{Q}\exp\left(-\frac{1}{2}(X - \mu_{h})\sum^{-1}(X - \mu_{h})^{\mathrm{T}}\right)}{\mathrm{P}\left[X \mid noise\right]}$	$resp = \left\{ \begin{array}{l} 0, \text{ if } \hat{x} < criterion \\ 1, \text{ if } \hat{x} > criterion \end{array} \right\}$ *criterion* is fit separately for each group
Associative heuristic (mixture condition)	$\hat{x} = \left\{ \begin{array}{l} 1, \text{ if } \max(X) \geq k \\ 0, \text{ if } \max(X) < k \end{array} \right\}$ k = 7, 8, 9, or 10, and is fit separately for each group	$resp = \hat{x}$

n = group size; $\hat{x}$ = group decision variable; x_j = rating of the j*th* group-member; $X = [x_1, x_2 \ldots x_n]$;
$\boldsymbol{\Sigma}^{-1}$ = inverse of the covariance matrix; superscript T refers to transpose; μ = mean rating row vector = $[u_1, u_2 \ldots u_n]$;
μ_{SIGNAL} = mean rating row vector for all signal trials; μ_{NOISE} = mean rating row vector for all noise trials;
$\mu_{\text{j-Strong}}$ = mean rating row vector for all trials that the j*th* group-member received the strong signal;
Q = normalization constant of the multivariate probability distribution = $\sqrt{|\Sigma|(2\pi)^n}$.

Results

Theoretical Simulations: Different information environments require different collective integration rules

To illustrate how different collective integration rules perform across diverse information environments, we implemented theoretical simulations using Gaussian random variables and evaluated the performance accuracy achieved by different integration rules: majority, averaging, weighted linear, and optimal Bayesian (see Appendix B for simulation details). The Gaussian distributions obeyed the statistical distribution of information across individuals and time in our experiments, but we explored many different signal strengths. For the equal signal-strength situations, the three individual decision variables (one per simulated observer) were sampled from distributions with unit standard deviations and with equal means (0 for noise trials; SNR for signal trials). We explored the outcome of using different signal strengths by systematically varying the signal to noise ratio (SNR). For the mixture situations, for noise trials the individual decision variables were again sampled from zero-mean distributions. For signal trials, one of the decision variables was sampled from a distribution centered on SNR_{strong}, while the other two decision variables were sampled from distributions centered on SNR_{weak}. Critically, the identity of the individual receiving the strong signal is randomized and unknown for each trial. We explored the outcome of using different strengths for the strong signal by systematically varying SNR_{strong} (for simplicity we kept SNR_{weak} constant throughout the simulations that we show in Figure 3B).

The simulation results in Figure 3A show that in *equal strength situations* the majority rule fares well even compared to the optimal (linear) rule. Conversely, the simulation results in Figure 3B show that in *mixture situations* the majority rule fares poorly compared to the averaging rule and the weighed linear combination rule, and extremely poorly compared to the optimal (non-linear) rule. The optimal Bayesian rule for the mixture situations is non-linear as it sums likelihoods across all mutually exclusive sets of signal-strength assignments to group members that are possible during signal trials (the three possible mutually exclusive sets are: (i) weak, weak, strong; (ii) weak, strong, weak; and (iii) strong, weak, weak). The exponential likelihood calculation is an accelerating non-linearity that effectively amplifies any one of the decisions variables arising from the three simulated individuals that attains a high value.

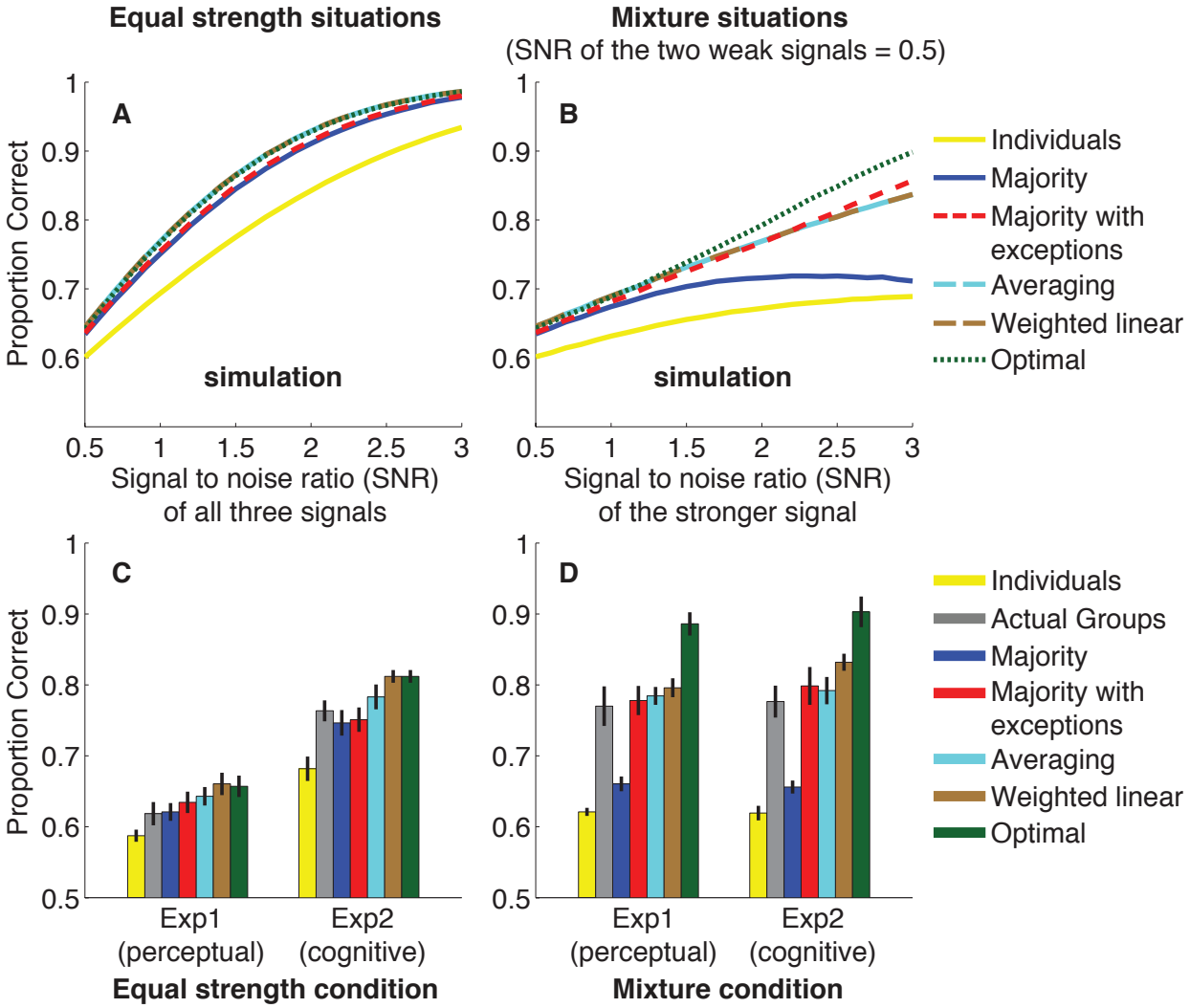

Figure 3. Proportion of trials that the decision outcome is correct for individuals, groups, and various group decision rules (group size = 3) in a yes/no signal detection task.
A. Theoretical simulations showing that when group members receive signals of equal strength there is little difference in performance between different group decision rules (see Appendix B for simulation details). **B.** Theoretical simulations showing that when on each signal trial a different random member of the group receives a stronger signal than the other two members, there is a wide difference in performance between different group decision rules. **C.** During the equal strength condition, when group members received signals of equal strength, there was little difference in performance between different group decision rules (error bars mark ±SEM). **D.** During the mixture condition, when on each signal trial a different random member of the group received a strong signal while the other two members received weak signals, there was a wide difference in performance between different group decision rules. Notice that while the majority rule (blue bars) outperforms the average individual accuracy of the group (yellow bars), it performs significantly worse than the accuracy of participants' actual group decisions (grey bars). This indicates that groups did not employ the majority rule during the mixture condition.

Benefit of group decisions relative to individual decisions

Consistent with previous studies (Sorkin et al., 2001), and as shown in Figure 3C-D, the overall accuracy (i.e., Proportion Correct) of human group decisions (grey bars) was significantly greater ($p < .01$) than the overall accuracy of individual decisions (yellow bars) for all conditions and tasks. The overall accuracy of group decisions, however, was not greater than the overall accuracy of the best performing member of each group for all conditions[2]. In Exp1 (perceptual task), the overall accuracy of group decisions during the *equal strength condition* (.62) was not significantly greater than that of the best performing member of each group (.63), $t(9)=1.76$, $p > .05$. On the other hand, the overall accuracy of group decisions during the *mixture condition* (.77) was significantly greater than that of the best performing member of each group (.64), $t(9)=4.64$, $p < .01$. Similarly in Exp2 (cognitive task), the overall accuracy of group decisions during the *equal strength condition* (.76) was not significantly greater than that of the best performing member of each group (.76), $t(9)=0.22$, $p > .05$. On the other hand, the overall accuracy of group decisions during the *mixture condition* (.78) was significantly greater than that of the best performing member of each group (.66), $t(9)=5.24$, $p < .01$.

Collective integration rules applied to participants' individual ratings

To assess the relative effectiveness of different collective decision rules, we quantified the performance accuracy of each rule applied to participants' personal confidence ratings and compared them to one another (i.e., Proportion Correct). Although one might expect that achieved accuracies for the various rules would be similar to those in the theoretical simulations, this does not necessarily have to be the case. Our simulations assume equal variance Gaussian distributions, continuous decision variables, and equal index of detectability (d') for each observer. If the actual observer ratings arise from internal decision variables that depart from Gaussian or have unequal variance, then the relative accuracies of the various group decision rules might be different. In addition, individual detection abilities differ across observers and that might impact the relationship across different group decision rules.

---

[2] For the equal strength condition, 4 out of 10 groups in Exp1 and 7 out of 10 groups in Exp2 attained a higher group-decision accuracy than the accuracy of the group's best performing member. For the mixture condition, 9 out of 10 groups in Exp1 and 10 out of 10 groups in Exp2 attained a higher group-decision accuracy than the accuracy of the group's best performing member.

The bottom panels of Figure 3 show the accuracies achieved by the various collective integration rules applied on a trial-by-trial basis to the observer ratings from the experiments. Consistent with the theoretical simulations in Figure 3A, Figure 3C shows that during the equal strength condition there is little difference in performance between the majority rule, the optimal (linear) rule, and other rules. Moreover, and consistent with the theoretical simulations in Figure 3B, Figure 3D shows that during the mixture condition the majority rule performs significantly worse than all the other rules, while the optimal (non-linear) rule performs significantly better than all the other rules.

A first comparison of interest is the overall accuracy of each collective integration rule compared to the actual group performance attained by humans. Arguably, if human groups achieve statistically significant higher accuracy than that achieved by a collective integration rule, it suggests that humans are not adopting that rule. As shown in Figure 3D, participants' actual group decisions during the mixture condition were significantly more accurate than the majority rule in both Exp1 ($t(9)=4.55$, $p < .01$) and Exp2 ($t(9)=6.06$, $p < .01$). This indicates that participants did not use the majority rule during the mixture condition.

Humans adapt their collective integration rules

To evaluate which collective decision rule the groups were actually adopting, we assessed how well different rules account for the participants' actual group decisions on a trial-by-trial basis (i.e., Choice Probability). During the equal strength condition, participants' group decisions were most consistent with the majority rule. The choice probability of the majority rule (.96) is significantly greater than the next closest competitor, which is the averaging rule (.9), in both Exp1 ($t(9)=3.78$, $p < .01$) and Exp2 ($t(9)=3.39$, $p < .01$). Note that all models *except* for the majority rule have one fitting parameter: either a group decision criterion (D_C) for the averaging, weighted linear, and the Bayesian models, or the individual rating value (k) that is considered a high enough endorsement of signal presence for the majority with exceptions model; see Table 1. Hence, the majority rule accounts best for the behavioral data during the equal strength condition despite not having a fitting parameter.

Conversely, during the mixture condition, the choice probability of the majority rule (.86) is no better than most other competing rules such as averaging (.9) and weighted linear combination (.9). Moreover, as shown in Figure 4, the choice probability of the majority rule

actually *decreases* over the course of the mixture condition in both Exp1 ($F(4,36)=2.91$, $p < .05$) and Exp2 ($F(4,36)=10.38$, $p < .01$). While the majority rule, as mentioned above, does *not* have a fitting parameter like the other rules, the significant downward trend of its choice probability indicates a gradual abandonment of the majority rule during the mixture condition, consistent with the hypothesis that humans adapt to the mixture situation and adopt some other integration rule that leads to better performance.

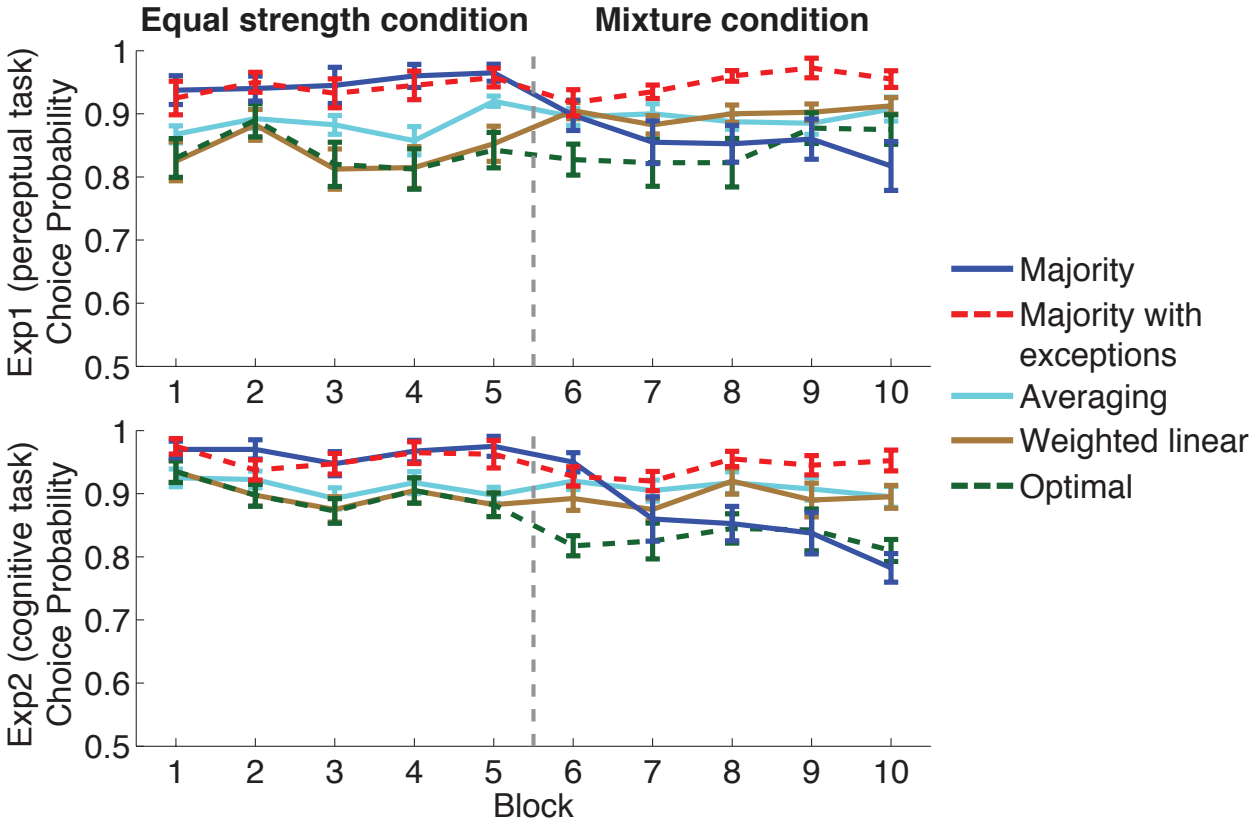

Figure 4. Choice probability. The choice probabilities of various group decision rules are shown in non-overlapping blocks of 40 trials (error bars mark ±SEM). During the equal strength condition, participants' group decisions are best accounted for by the majority rule (see Figure 5 and text regarding the majority with exceptions rule). During the mixture condition, participants' group decisions are best accounted for by the majority with exceptions rule.

To identify what group decision algorithm participants used during the mixture condition, we explored several different potential rules that groups may have adopted including averaging, weighted linear combination, the non-linear optimal Bayesian, and others. We found that participants' group decisions were most consistent with a heuristic of following the majority opinion unless someone was highly confident that it was a signal trial, in which case they followed that highly confident minority opinion. This heuristic (*majority with exceptions*) manages to avoid many pitfalls of the majority rule while remaining computationally simple. The choice probability for the majority with exceptions rule during the mixture condition (.94) is significantly greater than the next closest competitor, which is the averaging rule (.9), in both Exp1 ($t(9)=5.94$, $p < .01$) and Exp2 ($t(9)=3.02$, $p < .05$). Furthermore, as shown in Figure 4, the choice probability of this rule actually *increases* over the course of the mixture condition in Exp1 ($F(4,36)=2.92$, $p < .05$), though not significantly in Exp2 ($F(4,36)=1.14$, $p > .05$).

Majority with exceptions accounts best for participants' group decisions during the mixture condition

As mentioned above and detailed in Appendix A, all models (except for the majority rule) have a fitting parameter: a group decision criterion (D_C) for the averaging, weighted linear, and Bayesian models; and the individual rating value (k) that is considered a high enough endorsement of signal presence for the majority with exceptions model. We showed above and in Figure 4 that if the free parameter for each model is fit using all of the mixture condition trials, the choice probability of the majority with exceptions rule is significantly greater than all other rules. Is it possible, however, that averaging, weighted linear and/or optimal Bayesian have higher choice probability than majority with exceptions during the mixture condition if we allow the fitting parameter for each model to change from block to block?

Our modeling indicates that even if the free parameter for each model is fit block-by-block during the mixture condition, the other rules still fall short of the majority with exceptions rule in accounting for participants' group decisions during the mixture condition. The choice probability for the majority with exceptions rule with a shifting parameter k during the mixture condition (.95) is still significantly greater than the next closest competitor with a shifting parameter D_C, which is now the weighted linear rule (.92), in both Exp1 ($t(9)=4.97$, $p < .01$) and Exp2 ($t(9)=2.53$, $p < .05$). Furthermore, the choice probability of the majority with exceptions

rule with a shifting parameter k still *increases* over the course of the mixture condition in Exp1 ($F(4,36)=2.67$, $p < .05$), though not significantly in Exp2 ($F(4,36)=1.09$, $p > .05$).

An additional heuristic rule we evaluated for the mixture condition is one in which observers learn associatively through feedback to not only respond "yes" as a group when one or more group-members endorse signal presence with high confidence, but to also (unlike the majority with exceptions rule) respond "no" as a group in the *absence* of a highly confident opinion that it was a signal trial, even if all group-members individually endorse signal presence but with low confidence. The performance accuracy (i.e., Proportion Correct) of this *associative heuristic* rule during the mixture condition (.88) is very high and approximates that of the optimal non-linear rule (.89). However, our results indicate that participants did not adopt this highly effective associative heuristic rule during the mixture condition, because its choice probability (.81) is significantly lower than that of the majority with exceptions rule (.94) in both Exp1 ($t(9)=6.1$, $p < .01$) and Exp2 ($t(9)=5.86$, $p < .05$).

Analysis of conflict trials

Having identified the majority with exceptions rule to best account for participants' group decisions during the mixture condition, it is important to test how it fares during the equal strength condition. The choice probability of the majority with exceptions rule during the equal strength condition (.95) is very high and approximates that of the majority rule (.96). This raises the possibility that there was no adaptation at all during the mixture condition and that participants were perhaps using the majority with exceptions rule from the beginning of the experiments. To test for this possibility, we identified all trials across all groups where the majority with exceptions rule is in conflict with the majority rule (i.e., all trials where one member was highly confident that it was a signal trial while the two other members thought that it was a noise trial regardless of how confident they were). We found that while the percentage of trials with conflict between the two integration rules is relatively high during the mixture condition (30.6% of mixture condition trials were conflict trials), it is very low during the equal strength condition (6.9% of equal strength condition trials were conflict trials). This low percentage explains the close correspondence in overall choice probabilities between the majority rule and the majority with exceptions rule during the equal strength condition. However we further explored the relationship between the integration rules by isolating and computing the

proportion of conflict trials that groups followed the highly confident minority opinion rather than the majority opinion and tested whether the proportion is greater during the mixture condition relative to the equal strength condition. In addition, and given that there was a relatively high number of conflict trials during the mixture condition, we tested whether the proportion of conflict trials that are accounted for by the majority with exceptions rule changes over the course of the mixture condition.

Overall, there was a significant increase halfway through the experiments in the proportion of conflict trials that groups followed the highly confident minority opinion that said "yes" rather than the majority opinion that said "no". In Exp1, the majority with exceptions rule accounts for .41 of conflict trials (34 out of 83) in the equal strength condition and for .81 of conflict trials (237 out of 291) in the mixture condition, which constitutes a significant increase ($p < .02$; bootstrap resampling (Efron & Tibshirani, 1993)). In Exp2, the majority with exceptions rule accounts for .35 of conflict trials (19 out of 55) in the equal strength condition and for .76 of conflict trials (244 out of 321) in the mixture condition, which again constitutes a significant increase ($p < 2.2\text{x}10^{-4}$; bootstrap resampling).

In Figure 5 we broke the mixture condition into five non-overlapping blocks and show the proportion of conflict trials that groups followed the highly confident minority opinion rather than the majority opinion as the experiments progressed. The upward trend in favor of the high confidence minority opinion during the mixture condition (see the best fitting trend lines that are shown in Figure 5) is in good agreement with the gradual abandonment of the majority rule discussed above (and shown in Figure 4). The inserts in Figure 5 show a histogram of the bootstrap estimates for the slope of this upward trend during the mixture condition. The slope of the best fitting trend line during the mixture condition is significantly greater than zero in both Exp1 (slope=.082, $p < 4.5\text{x}10^{-5}$; bootstrap resampling) and Exp2 (slope=.126, $p < 1.0\text{x}10^{-6}$; bootstrap resampling). These results indicate that groups dynamically changed their integration rule to increasingly follow minority opinions endorsing signal presence with high confidence.

We also add that if we allow a shifting parameter k (so as to maximize the choice probability of the majority with exceptions rule block-by-block during the mixture condition), the results still show an upward (though shallower) trend in favor of the high confidence minority opinion during the conflict trials of the mixture condition. Concretely, even if the exception parameter k is fit block-by-block during the mixture condition, the slope of the best

fitting trend line is still significantly greater than zero in both Exp1 (slope=.06, $p < 1.5\text{x}10^{-3}$; bootstrap resampling) and Exp2 (slope=.10, $p < 1.6\text{x}10^{-4}$; bootstrap resampling). This indicates a gradual abandonment of the majority rule in favor of the majority with exceptions rule that goes *beyond* any fine-tuning of the exception parameter *k* during the mixture condition.

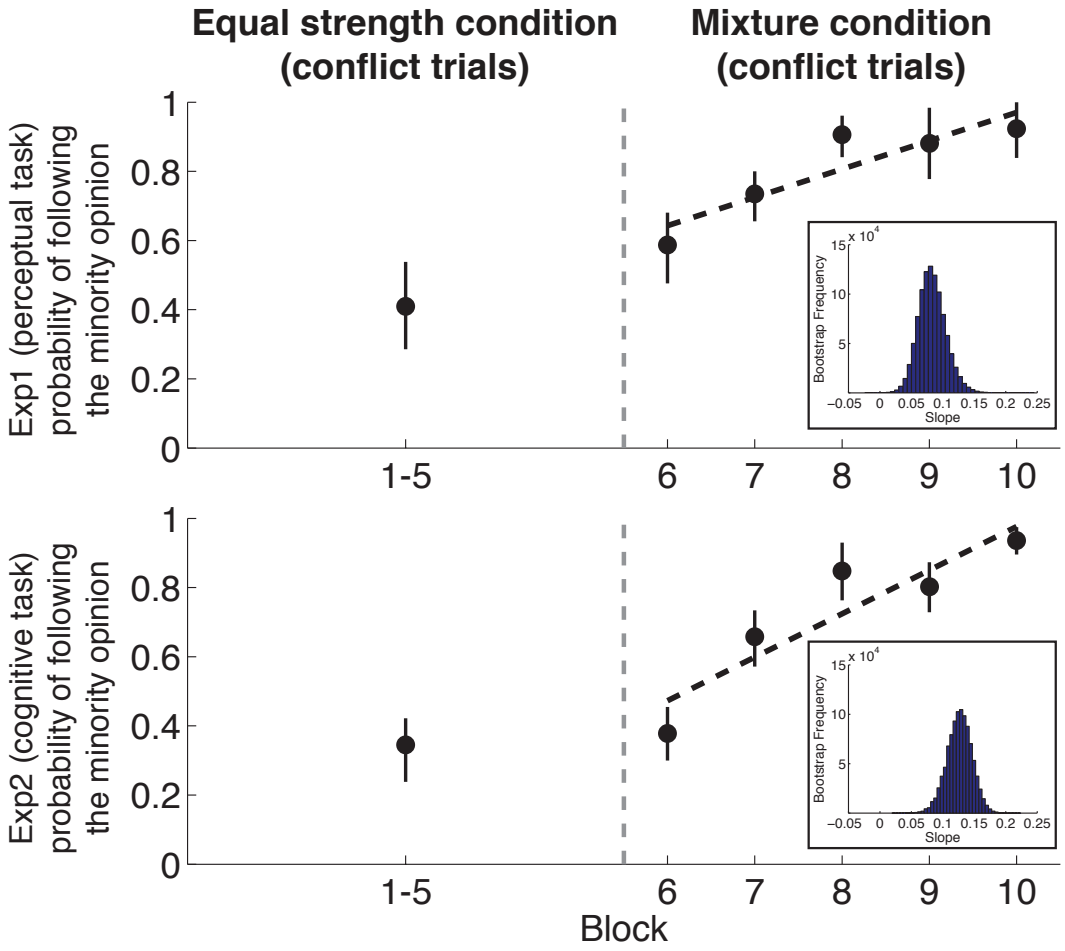

Figure 5. Conflict trials. We isolated all trials across all groups were the prediction of the majority with exceptions rule is in conflict with the majority rule. Data points show the proportion of conflict trials that groups followed the highly confident *minority opinion* rather than the majority opinion as the experiments progressed (error bars mark bootstrap 68.27% confidence intervals to be equivalent to the percentile of ±SEM of a normal distribution). The dashed diagonal lines mark the best fitting trend line during the mixture condition, and the inserts show the histogram of the bootstrap estimates for that slope. A positive slope indicates an upward trend during the mixture condition in the proportion of conflict trials that groups followed the highly confident minority opinion endorsing *signal presence* rather than the majority opinion endorsing *signal absence*.

Group decision accuracy is correlated with propensity to abandon the majority rule

Figure 6 shows that while almost all groups eventually abandoned the majority rule, some adapted to the mixture of distributions better and/or quicker than others. We expect that groups that were less prone to abandon the majority rule during the mixture condition should attain lower collective decision accuracies during the mixture condition. The top half of Figure 7 shows that, indeed, there is a strong *negative correlation* between the choice probability of the majority rule and actual group performance in both Exp1 ($r(8)=-.91$, $p < .01$) and Exp2 ($r(8)=-.82$, $p < .01$). Conversely, the more a group's decisions resemble the optimal non-linear rule the greater its decision accuracy, as evidenced by the strong *positive correlation* (bottom half of Figure 7) between the choice probability of the optimal non-linear rule and actual group performance in both Exp1 ($r(8)=.95$, $p < .01$) and Exp2 ($r(8)=.78$, $p < .01$). But this does not mean that high performing groups were actually implementing the optimal non-linear rule during the mixture condition, because its choice probability (.84) is lower than that of the majority with exceptions rule (.94) for every single group in both Exp1 ($t(9)=4.25$, $p < .01$) and Exp2 ($t(9)=6.62$, $p < .01$).

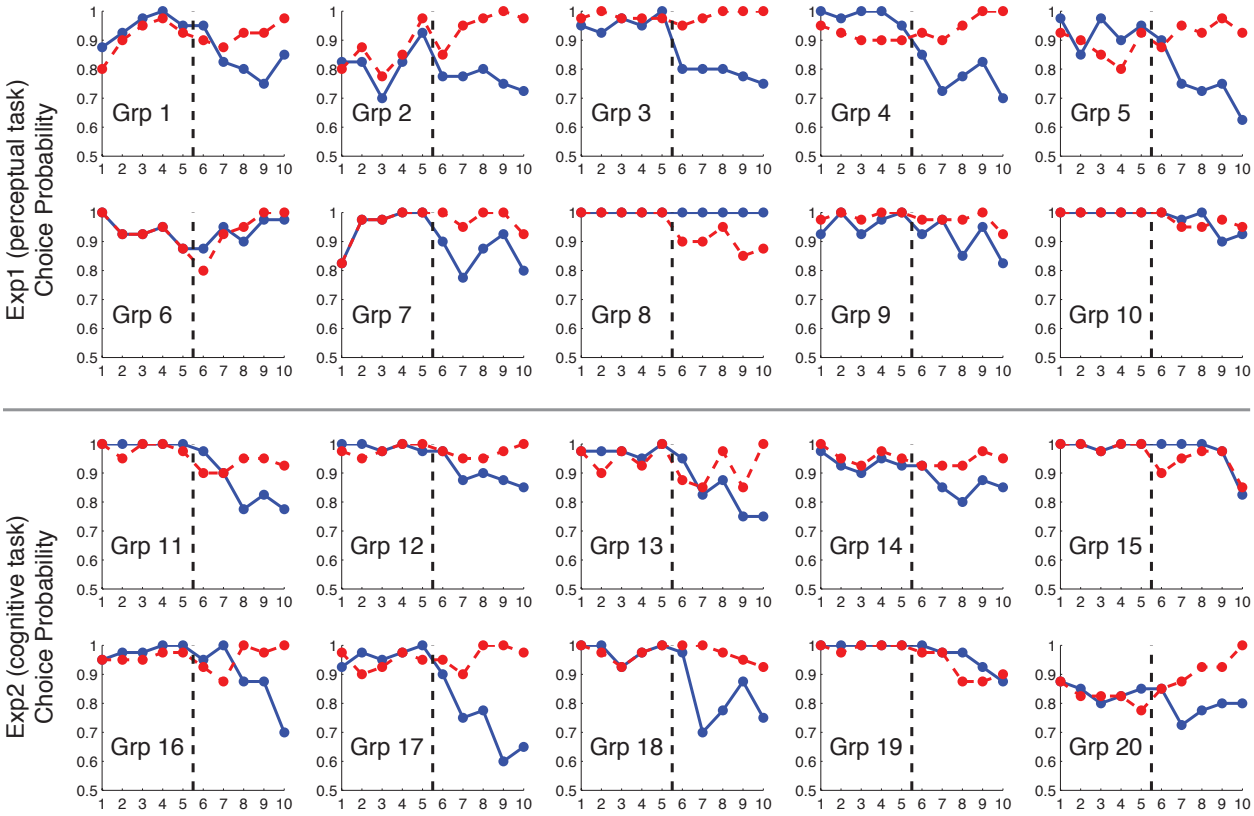

Figure 6. Choice probabilities of the Majority rule (blue) and the Majority with exceptions rule (red) for each group. The choice probabilities of the two rules are shown in non-overlapping blocks of 40 trials. During the equal strength condition (blocks 1-5), participants' group decisions are well accounted for by the majority rule. During the mixture condition (blocks 6-10), participants' group decisions are best accounted for by the majority with exceptions rule. *Note that some groups adapted to the mixture situation better and/or quicker than others.* Remarkably, Group 8 never adapted to the mixture situation as their group decisions were always accounted for by the majority rule.

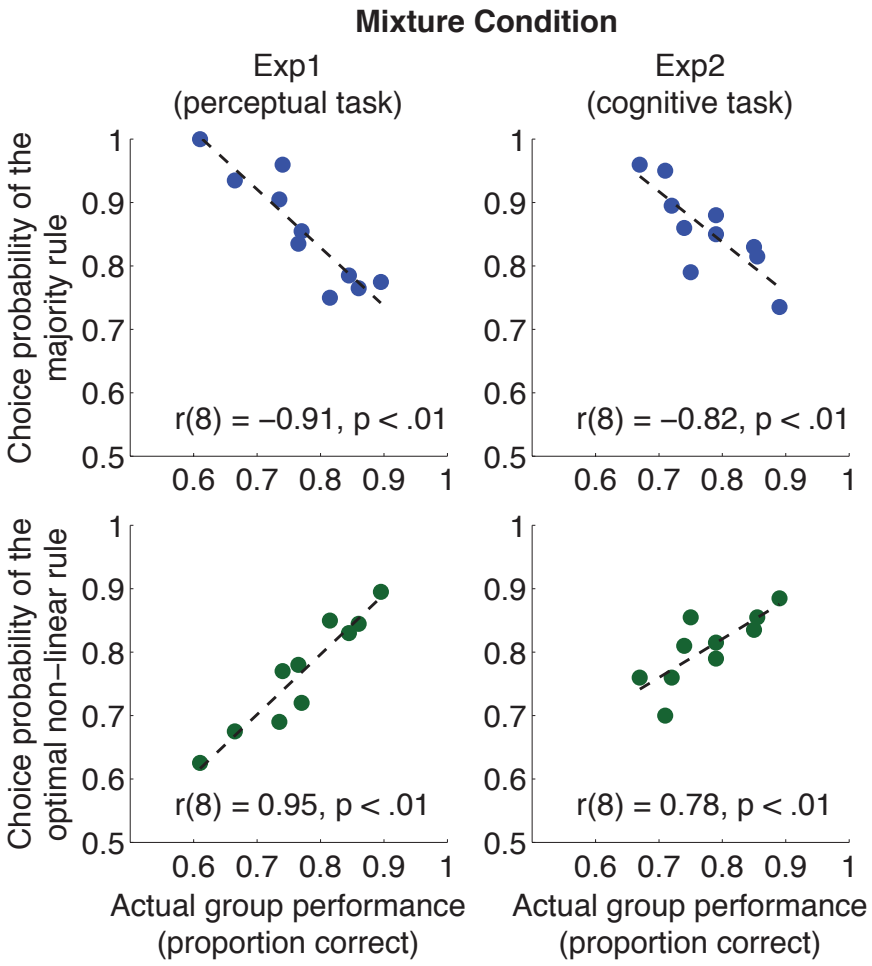

Figure 7. Correlation. Actual performance of each group during the mixture condition (proportion of mixture condition trials that the group decision is correct) is plotted against the respective choice probabilities of the majority rule (top panels) and the optimal Bayesian non-linear rule (bottom panels).

In what way do humans diverge from the optimal Bayesian non-linear rule in the mixture condition?

Although the results show that humans flexibly adapted their group decision rules to changes in the statistical distribution of information across individuals and time, human performance is significantly lower than the optimal Bayesian non-linear rule. Here we explore the types of trials for which human group decisions systematically differ from the optimal non-linear rule in the mixture condition. The color scales in Figure 8 represent the proportion of trials that the group response is "no" (left column) or "yes" (right column) for the optimal Bayesian non-linear rule (top panels), the majority rule (middle panels), and participants' actual group responses (bottom panels), as a function of (i) the number of individuals in the group choosing "no" (left column) or "yes" (right column) and (ii) the value of the lowest individual rating (left column) or the highest individual rating (right column). Qualitatively the results of this analysis are the same in both experiments, so for simplicity we show data collapsed across both experiments. The numbers in the cells show the total number of each type of trial across all 20 groups[3].

The majority rule (middle panels) is straightforward: it always responds "no" when two or more individuals endorse *signal absence* irrespective of the value of the lowest individual rating, and it always responds "yes" when two or more individuals endorse *signal presence* irrespective of the value of the highest individual rating. Hence, the majority rule is exclusively sensitive to the number of individuals endorsing signal presence/absence, and completely *insensitive* to the highest/lowest individual rating.

The optimal Bayesian non-linear rule (top panels) and humans (bottom panels) both depart from the majority rule but in different ways. The top right panel shows that the optimal non-linear rule is insensitive to the number of individuals endorsing signal presence; instead, its decisions are determined by the value of the highest individual rating. Of all the panels in Figure 8, this is the most striking departure from the majority rule. Compare this to the top left panel where the optimal non-linear rule is more sensitive to the number of individuals endorsing signal absence (similar to the majority rule) rather than the value of the lowest individual rating. Hence

[3] Note that many trials are repeated in both columns because when two group-members are endorsing signal absence (left column) one is endorsing signal presence (right column), and when one is endorsing signal absence two are endorsing signal presence.

there is an asymmetry in that the optimal Bayesian non-linear model is very sensitive to ratings that are very high but not to those that are very low.

As for humans, the bottom right panel shows that when only one individual (i.e., a minority opinion) endorses signal presence, humans are very sensitive to the value of that individual's rating just like the optimal non-linear rule. But, when two or more individuals (i.e., a majority opinion) endorse signal presence, humans, unlike the optimal non-linear rule, are very *insensitive* to the value of the highest individual rating, because the group response is almost always "yes" (similar to the majority rule). This indicates that human group decisions resemble the optimal Bayesian non-linear rule in some circumstances (when a *minority* endorses signal presence) but not others (when a *majority* endorse signal presence). To follow we explore this directly.

The color scales in the top panels of Figure 9 represent the proportion of trials that human group decisions *disagree* with the decision of the optimal Bayesian non-linear rule for each type of trial. The color scales in the bottom panels represent the *increase in accuracy* that the optimal Bayesian non-linear rule achieves above and beyond the accuracy of human group decisions for each type of trial. We expect that higher disagreement levels should lead to higher increases in accuracy for the optimal Bayesian non-linear rule relative to humans.

The top left panel of Figure 9 shows that participants' actual group responses are in good agreement with the optimal Bayesian non-linear rule as a function of how low the lowest individual rating was and the number of individuals endorsing *signal absence*. The bottom left panel shows that this low level of disagreement in those types of trials leads to only modest increases in accuracy for the optimal non-linear rule relative to humans. The top right panel shows that the level of disagreement between humans and optimal varies as a function of how high the highest individual rating was and the number of individuals endorsing *signal presence*. Most cells show relatively low levels of disagreement, while some cells show very high levels of disagreement, particularly in the four cells when two or three group-members endorse signal presence and the highest individual rating is 6 or 7 (i.e., when a majority endorse signal presence but with low confidence). The level of disagreement between the optimal non-linear rule and human group decisions for the trials in those four cells (.86) is significantly greater than that in the next four highest cells with highest disagreement (.45) in both Exp1 ($p < 9\text{x}10^{-6}$; bootstrap resampling) and Exp2 ($p < 2\text{x}10^{-6}$; bootstrap resampling).

Similarly, the bottom right panel shows that the increase in accuracy for the optimal Bayesian non-linear rule relative to humans is modest in most cells, but quite high in the same four cells when a majority of group-members endorse signal presence but with low confidence. The increase in accuracy for the optimal non-linear rule relative to humans for the trials in those four cells (.57) is significantly greater than that in the next four highest cells (.2) in both Exp1 ($p < .02$; bootstrap resampling) and Exp2 ($p < 1.1\text{x}10^{-3}$; bootstrap resampling). In those trials when a majority of group-members endorse signal presence but with low confidence, the optimal Bayesian non-linear model assumes that the signal is *absent* and it responds "no", while human groups assume that the signal is *present* and they collectively respond "yes" (cf. the right column of Figure 8).

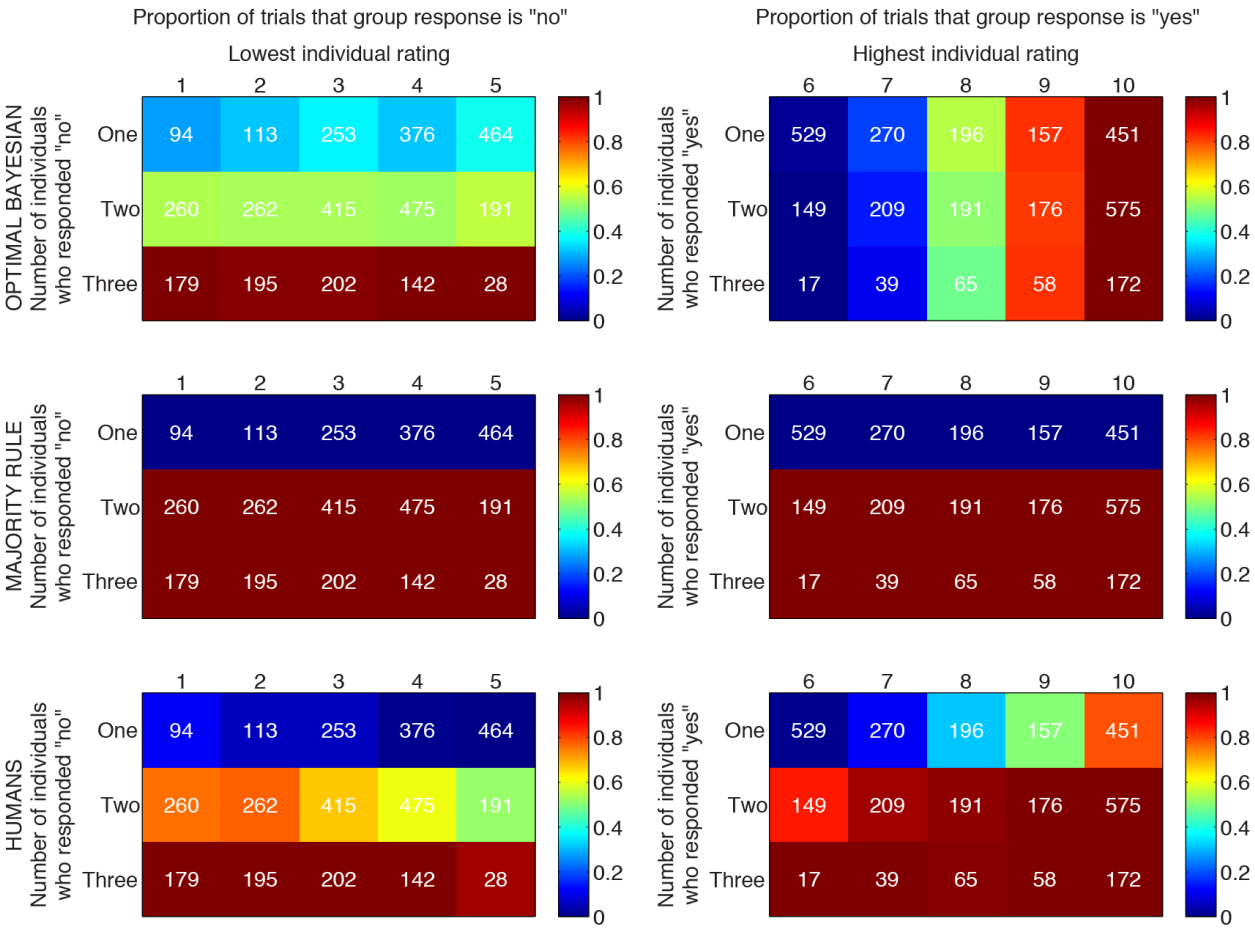

Figure 8. Trial types during the mixture condition. The numbers in the cells show the total number of each type of trial across all 20 groups during the mixture condition. **Left column.** Proportion of trials that the group response is "no" for the optimal Bayesian non-linear rule (top panel), the majority rule (middle panel), and participants' actual group responses (bottom panel) as a function of (i) how many group-members endorsed *signal absence* by individually responding "no" and (ii) the value of the *lowest* individual rating. **Right column.** Proportion of trials that group response is "yes" for the optimal Bayesian non-linear rule (top panel), the majority rule (middle panel), and participants' actual group responses (bottom panel) as a function of (i) how many group-members endorsed *signal presence* by individually responding "yes" and (ii) the value of the *highest* individual rating.

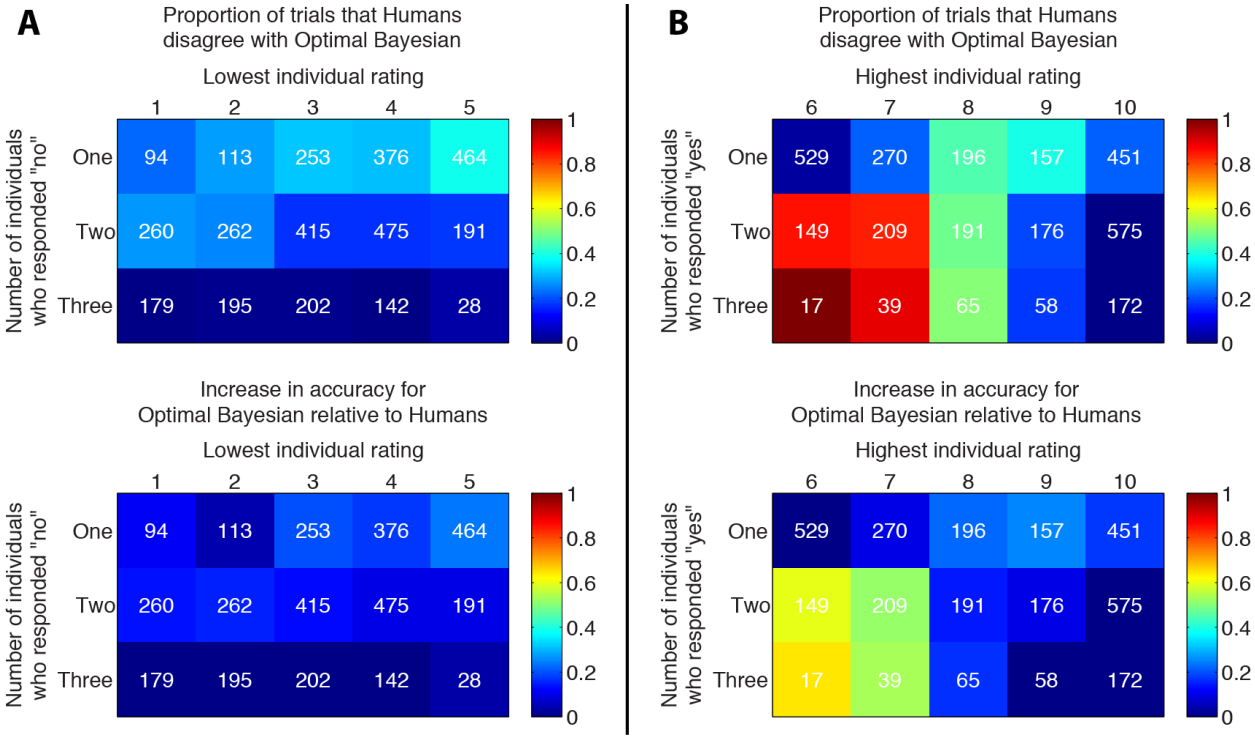

Figure 9. Human comparison to the optimal Bayesian non-linear model during the mixture condition. The numbers in the cells show the total number of each type of trial across all 20 groups during the mixture condition. **A.** *Top panel.* Proportion of trials that participants' actual group responses are in disagreement with the optimal Bayesian non-linear model as a function of (i) how many group-members individually responded "no" and (ii) the value of the lowest individual rating. *Bottom panel.* Increase in accuracy that the optimal Bayesian non-linear rule achieves above and beyond the accuracy of human group decisions. **B.** *Top panel.* Proportion of trials that participants' actual group responses are in disagreement with the optimal Bayesian non-linear model as a function of (i) how many group-members individually responded "yes" and (ii) the value of the highest individual rating. *Bottom panel.* Increase in accuracy that the optimal Bayesian non-linear rule achieves above and beyond the accuracy of human group decisions.

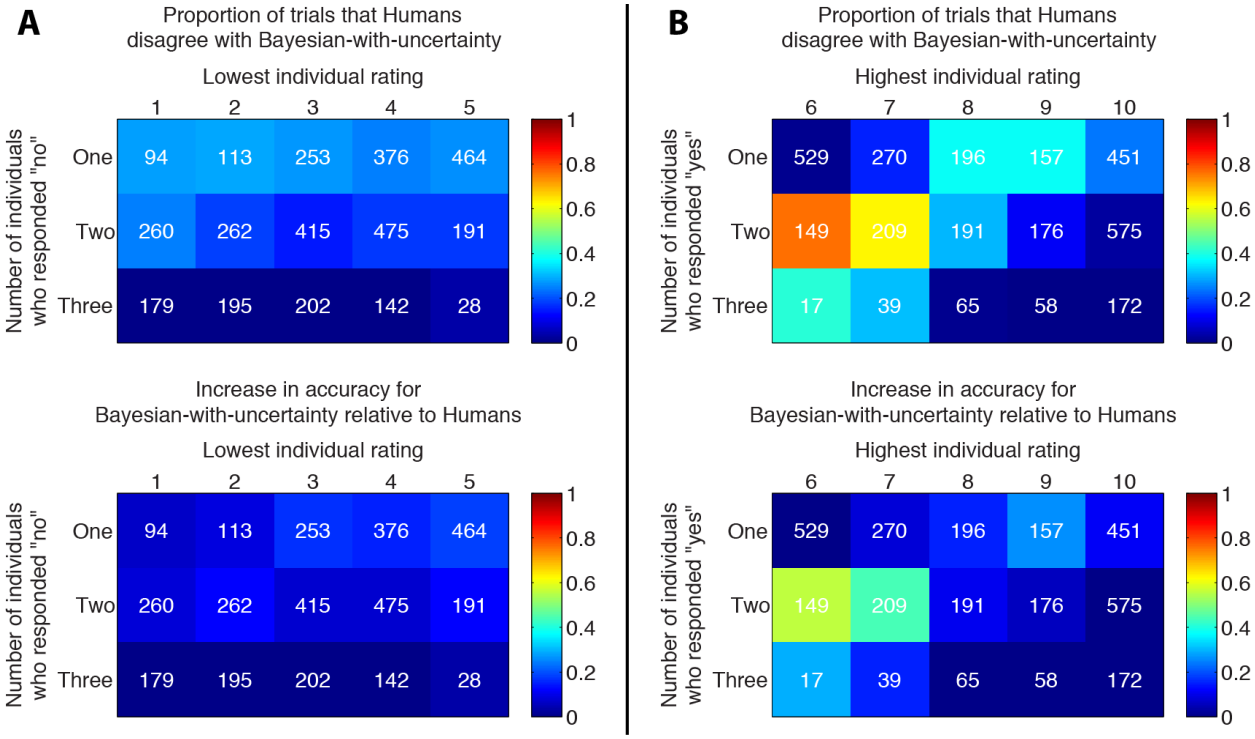

Figure 10. Human comparison to the Bayesian-with-uncertainty model during the mixture condition. This is the same as the previous figure, except that here we show results for the *Bayesian-with-uncertainty* model that considers many different combinations of signal-strength assignments to group members (as opposed to the optimal non-linear model in Figure 9 which is afforded knowledge that there are only three possible combinations). Notice the reduced level of disagreement in the top right panel between this model and humans (compared to the top right panel in Figure 9 for the optimal non-linear model), especially when two or more individuals endorse signal presence and the highest individual rating is only 6 or 7 (i.e., when a majority of group-members endorse signal presence but with low confidence).

Retaining elements of the majority rule (to combat uncertainty about signal-strength)

In the previous section we showed that the greatest disagreement between humans and optimal is when a majority of group-members endorse signal presence but with low confidence. Unlike the optimal non-linear model, humans did not infer that the group response should be "no" in the *absence* of a high confidence endorsement of signal presence. One possibility we consider is that humans might retain elements of the majority rule as a form of robustness (Huber, 1981) to further changes in the statistical distribution of information across individuals and time. To explore this possibility, we implemented a non-linear Bayesian model that considers many different combinations of signal-strength assignments to group members, and we assessed whether it results in greater agreement with human group decisions during the mixture condition.

To introduce uncertainty regarding signal-strengths, the *Bayesian-with-uncertainty* model considers many different sets of signal-strength assignments using 15 equally spaced steps that range from each group-member's mean rating when he/she received a strong signal, down until (but not including) his/her mean rating during noise trials. These steps were recomputed every single trial using a leave-one-out procedure (i.e., leaving out the rating of the current trial). Hence, this non-linear Bayesian model with signal strength uncertainty considers 3375 mutually exclusive sets of signal-strength assignments to group members ($15^3 = 3375$), compared to just the three sets that the optimal Bayesian non-linear model considers.

Overall, the *Bayesian-with-uncertainty* model resembles human behavior during the mixture condition more so than the optimal non-linear model does. Across all mixture condition trials, while the choice probability of the Bayesian-with-uncertainty model (.89) is not as high as that of the majority with exceptions rule (.94), it is significantly greater than that of the optimal non-linear model (.84) in both Exp1 ($t(9)=3.29$, $p < .01$) and Exp2 ($t(9)=9.37$, $p < .01$). When looking more closely at specific types of trials, the top right panel of Figure 10 shows that there is now less disagreement between this model and humans in the four cells when two or more individuals endorse signal presence and the highest individual rating is only 6 or 7 (compare to the same four cells in the top right panel of Figure 9 for the optimal non-linear rule). The level of disagreement with humans for the trials in those four cells for the Bayesian-with-uncertainty model (.65) is significantly lower than that for the optimal non-linear model (.86) in both Exp1 ($p < 1.3\text{x}10^{-3}$; bootstrap resampling) and Exp2 ($p < 7.7\text{x}10^{-5}$; bootstrap resampling).

Similarly, the bottom right panel of Figure 10 shows that there is now less benefit in accuracy for the Bayesian-with-uncertainty model relative to humans when two or more individuals endorse signal presence and the highest individual rating is only 6 or 7 (compare to the same four cells in the bottom right panel of Figure 9 for the optimal non-linear rule). The increase in accuracy relative to humans for the trials in those four cells for the Bayesian-with-uncertainty model (.46) is lower than that for the optimal non-linear model (.57), though not significantly in Exp1 (p = .09; bootstrap resampling) and significantly in Exp2 ($p < 3.9\text{x}10^{-3}$; bootstrap resampling). In those trials when a majority of group-members endorse signal presence but with low confidence, the *Bayesian-with-uncertainty* model does not respond "no" as often as the optimal non-linear model does, which leads to a lower increase in accuracy relative to humans for those types of trials (though not significantly in Exp1).

This analysis shows that increased uncertainty about signal strengths leads to a Bayesian model that agrees with the majority rule more often than the optimal non-linear model does, though not as often as humans do. One could argue then that in retaining elements of the majority rule, humans employed an algorithm (majority with exceptions) that reflects a human expectation that the signal strengths and their distribution across individuals might change again. Thus, the lack of complete adaptation to the particular mixture of distributions that they have encountered might allow robustness to future changes in the statistical distribution of information across individuals and time (Huber, 1981).

Discussion

1. Human flexible collective wisdom

We investigated how groups of humans reach collective decisions in response to changes in their information environment. We show that humans infer that different group-members might have access to dissimilar amounts of information at different times, and that they dynamically adapt their aggregation algorithms to improve the benefits of collective decision-making. Our finding is based on several converging results indicating a gradual abandonment of the majority rule during the mixture condition: (i) Humans achieve higher group-decision accuracies than the majority rule during the mixture condition but not during the equal strength condition (compare grey bars to blue bars in Figure 3C-D); (ii) The choice probability of the majority rule is lower than the choice probability of other rules (and especially the majority with exceptions rule) during the mixture condition but not during the equal strength condition; (iii) The choice probability of the majority rule *decreases* as the mixture condition progressed, while the choice probability of the majority with exceptions rule *increases* as the mixture condition progressed (see blue and red lines in blocks 6-10 in Figure 4); (iv) The majority with exceptions rule accounts for increasingly more *conflict trials* as the mixture condition progressed (see Figure 5); and our shifting parameter analysis shows that the gradual abandonment of the majority rule in favor of the majority with exceptions rule goes *beyond* any potential fine-tuning of the exception parameter k during the mixture condition.

This flexibility to employ different integration algorithms represents a departure from previous theories that conceptualize human collective decision-making using static decision rules. Although many studies from the fields of economic, behavioral and political science (Kalven & Zeisel, 1966; Kameda et al., 2003; Tindale, 1989), and vision science (Denkiewicz et al., 2013) demonstrate that human groups tend to employ a majority voting rule, one study found that when two observers make joint perceptual decisions (and the majority rule is not available (Denkiewicz et al., 2013)) they linearly adjust their relative weighting based on the approximate reliability of each individual (Bahrami et al., 2010), similar to the linear adjustment of relative weights in serial cue combination (Juni, Gureckis, & Maloney, 2012) and multisensory

integration within a single human (Ernst & Banks, 2002)[4]. Our results imply a more fundamental dynamic changing of the collective integration algorithm, from majority to majority with exceptions, in response to changes in the temporal distribution of information across individuals.

A possible alternative to our interpretation is that groups were using the same, single algorithm throughout the experiment while changing a parameter. The majority rule can be parameterized in our case as the majority with exceptions rule with $k = 11$ (so that the exception is never invoked). Hence, the apparent transition from the majority rule (during the equal strength condition) to majority with exceptions rule (during the mixture condition) can be interpreted as a change in the exception parameter k instead of a change of the actual integration algorithm. However, we note that many rules that are traditionally considered distinct can be framed and parameterized within a more general model. For example, even the majority and averaging rules in our case can be framed within a more general model (the trimmed mean rule), with a single parameter controlling any apparent transition from one rule to the other[5].

2. Groups abandon the majority rule because they infer that another rule performs better in their novel, complex information environment

In simple information environments, such as the equal strength condition, the majority rule fares well compared to other rules, and groups tend to rely on it when it is available because it is easy to implement and there is little accuracy to be gained by using some other rule instead. In complex information environments, such as the mixture condition, the majority rule fares poorly compared to other rules, and groups employ some other rule instead to increase the accuracy of their collective decisions.

---

[4] Note that Bahrami et al. (2010) show a suboptimal weighting inversely proportional to standard deviation ($1/\sigma$) for communicating dyads (cf. Ernst, 2010), while Ernst & Banks (2002) show an optimal weighting inversely proportional to variance ($1/\sigma^2$) for sensory cue integration within single humans.

[5] The free parameter in the more general model is the number of observations that are truncated for the trimmed mean. Each trial in our case consists of an odd number of individual ratings. If we sort the three ratings for each trial so that $x_1 \leq x_2 \leq x_3$, the yes/no response of the majority rule is identical to that of the trimmed mean rule if all ratings are removed except for the middle one x_2. Similarly, the yes/no response of the averaging rule is identical to that of the trimmed mean rule if no ratings at all are removed. Hence, the majority and averaging rules in our case can be framed and parameterized as variants of the trimmed mean rule, with different levels of truncation mapping onto different rules as follows: zero truncation is equivalent to the averaging rule, and extreme truncation is equivalent to the majority rule.

One might argue, at first glance, that our groups abandoned the majority rule during the second half of the experiments simply because they experienced performance loss associated with the majority rule when their information environment switched to a mixture of distributions. This *performance loss* hypothesis might explain the cognitive task (Exp2) where the performance accuracy of the majority rule is lower in the mixture condition (.66) than in the equal strength condition (.75). However, this hypothesis is inconsistent with our findings for the perceptual task (Exp1) where the performance accuracy of the majority rule is actually higher in the mixture condition (.66) than in the equal strength condition (.62). Instead, we hypothesize that our groups gradually abandoned the majority rule during the mixture condition because they inferred that they could increase the accuracy of their collective decisions if they employed some other rule instead.

3. Why do groups depart from the optimal Bayesian non-linear rule?

The human algorithm that emerges in the mixture condition (majority with exceptions) falls short of the optimal Bayesian non-linear algorithm. Aside from following minority opinions endorsing signal presence with high confidence, just like human groups do, the optimal non-linear rule also *rejects* majority (or unanimous) opinions endorsing signal presence if there is not a member amongst the majority advocating signal presence with high confidence. This leads the optimal non-linear rule to follow minority opinions endorsing *signal absence* much more often than human groups do.

Why do humans show this departure from the optimal Bayesian non-linear rule? The optimal non-linear model has full knowledge about the signal strengths that are possible and their statistical distribution across individuals and time (i.e., it knows that there are exactly three possible sets of signal-strength assignments to group members during signal trials). In contrast, the human participants were never instructed about the signal strengths or the statistical distributions. In addition, while the optimal model has knowledge that the statistical distributions would remain unchanged throughout the mixture condition, humans might assume and expect these to change at any time, especially given that they already experienced a change to their information environment halfway through the experiment. Thus we speculate that the human departures from the optimal non-linear rule might reflect the mistaken assumptions and uncertainties of the human participants.

To test this possibility, we investigated a Bayesian model that has uncertainty about the signal strengths that are possible and their statistical distribution across individuals and time. Such a model would be optimal for a continually changing information environment. Our results show that this *Bayesian-with-uncertainty* model better resembles the trial-by-trial decisions of human groups, and thus suggests that human departures from optimality in our tasks might reflect a combination of uncertainties about signal strength, and robustness to temporal changes in the distribution of information across group members.

Unlike our lab experiments, real world circumstances often provide contextual cues alerting the group that at any given time one person might have increased task-relevant information relative to the others. A group playing trivial pursuit knows that different people have expertise in different types of questions, and thus the group infers that they should follow different members' high confidence opinions at different times. A search party that disperses to find a missing child knows that any member of the search party could find strong evidence of the child's whereabouts (e.g., hearing their cry, or seeing their clothing), and thus the group intuitively understands that they should then refocus their rescue efforts to that particular location. Our experiments tried to minimize contextual cues by using synthetic perceptual (Exp1) and cognitive (Exp2) tasks; and yet, even in these cue-impoverished circumstances humans manage (albeit sub-optimally) to alter their collective decision rules.

Although our experiments highlight changes towards a majority with exceptions rule, it is likely that humans have access to a battery of aggregation rules and apply different distinct rules to different information environments in order to increase the benefits of collective decisions. The extent to which other species flexibly alter their collective integration algorithms in response to different information environments, or whether this flexibility might represent a hallmark of collective wisdom in organisms with highly evolved cognition remains to be demonstrated.

4. When do groups outperform the best individual in the group?

During the mixture condition, groups outperformed the best individual in the group; conversely, during the equal strength condition, half of the groups did not manage to outperform the best individual in the group (see footnote 2 for breakdown by condition and experiment). This pattern of results with triads (groups of three) stands in contrast to Bahrami et al. (2010) who found that dyads (groups of two) outperformed the best individual in the group when participants viewed

identical stimuli (experiment 1), but not when they viewed dissimilar stimuli with different levels of noise being added to each participant's stimulus in random order (experiment 2). To follow we consider each condition in turn.

With respect to the equal strength condition, it is important to note whether or not participants are partially correlated with each other due to external factors such as noise in the stimulus. Unlike the first experiment in Bahrami et al. (2010) that used perceptual stimuli without any external noise, our perceptual stimuli in Exp1 had additive luminance white noise, and our fictitious environmental measurements in Exp2 were partially correlated across the three computers (correlation = .2); see Experimental Methods. When employing a non-optimal integration algorithm like the majority rule or the averaging rule, the theoretical chances of the group outperforming the best individual goes down as the correlation across participants' judgments goes up due to identical external factors (this result was verified for groups of three via computer simulations with partially correlated random variables). Hence, our findings with participants who are partially correlated due to external factors do not necessarily contradict the first experiment in Bahrami et al. (2010) where participants are shown noiseless stimuli. Furthermore, we note that Denkiewicz et al. (2013) implemented the noiseless stimuli experiment in Bahrami et al. (2010) but with triads instead of dyads and they found, like us, that groups did not manage to outperform the best individual in the group. Similar to Denkiewicz et al. (2013), we speculate that social factors detrimental to group performance, such as social loafing (Latané, Williams, & Harkins, 1979) and social conformity (Asch, 1956), are more likely to be present with increased group size (cf. Ringlemann effect (Ingham, Levinger, Graves, & Peckham, 1974; Kravitz & Martin, 1986)), which could explain why triads do not manage to outperform the best individual in the group (Denkiewicz et al., 2013) while dyads do (Bahrami et al., 2010).

With respect to the mixture condition, when participants are shown signals with different strength, there is no contradiction between our findings and the second experiment in Bahrami et al. (2010) for the following simple reasons: Our analysis showing that the best individual does not beat the group considers *all* of the mixture condition trials that each individual encountered: 1/2 of the trials for each individual are noise trials, 1/6 are strong-signal trials, and 2/6 are weak-signal trials. On the other hand, the analysis in Bahrami et al. (2010) showing that the best individual beats the group only considers the trials in which the observer in question received a

strong signal (while the other observer received a weak signal) rather than all of the intermixed strong and weak signal trials for the observer in question. Thus, the apparent difference in findings between our study and Bahrami et al. (2010) reflects different comparisons. If in our study we compare each individual to the group but leave out the trials in which the individual in question received a weak signal, we find like Bahrami et al. (2010) that almost all individuals had either a higher discrimination ability (d') than the group (46 out of 60 individuals) or an equal d' (12 out of 60 individuals), with only two individuals having a lower d' than the group.

5. Limitations of current work

Our approach is to computationally model participants' group decisions based on their independent individual decisions, without controlling for social factors that might influence their individual or group decisions. We grouped participants together based on their schedule and availability without regard to their demographics or how motivated they were. Nevertheless, we recognize that these and many other factors and group dynamics (Lewin, 1947) could affect participants' individual and group decisions.

For example, while we tried to maintain independence between participants so that they would announce their *true* personal confidence ratings on each trial (see Group Decision section under General Methods), it is possible that being part of a group could lead to reduced cognitive effort (Petty, Harkins, Williams, & Latané, 1977) and negatively affect participants' individual performance, a phenomenon known as social loafing (Latané et al., 1979). With respect to group decisions, it is possible that social loafing (Latané et al., 1979) and/or pressure to conform to the rest of the group (cf. social conformity (Asch, 1956)) could lead individuals to avoid expressing their opinion too forcefully when discussing what the group decision should be, which could negatively affect group performance. For more on such factors see, amongst many other papers, the survey articles on group decision making in the Annual Review of Psychology (Kerr & Tindale, 2004; Levine & Moreland, 1990; McGrath & Kravitz, 1982).

While participants in our perceptual task (Exp1) were not faced with any spatial uncertainty (the signal was always placed at the center of the screen), future studies can randomize the location of the signal (to turn the experiment into a visual search task) and explore if, how, why, and when participants employ some division of cognitive labor to improve their collective performance, which could be mediated through social collaboration mechanisms

(Brennan & Enns, 2014). If allowed to communicate freely with each other, participants could potentially devise a "divide and conquer" strategy with each participant searching for the signal in a different part of the image, and the benefits of collective decision-making might exceed those measured in the current study.

Finally, the current study explores the adaptability of groups to collectively integrate information provided to different individuals. Of possible interest for futures studies is to investigate whether individual decision-makers adapt their algorithm to integrate across multiple observations of other individuals (Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Soll & Larrick, 2009), or even of their own (Herzog & Hertwig, 2009), based on the distribution of information across observations to maximize the wisdom of crowds (Galton, 1907; Surowiecki, 2005), or even the wisdom of the crowd within (Herzog & Hertwig, 2014; Vul & Pashler, 2008).

Acknowledgments

We thank Craig Abbey for many helpful discussions. This project was supported by a grant from the Intelligence Community Postdoctoral Research Fellowship Program through funding from the Office of the Director of National Intelligence, and by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. http://doi.org/http://dx.doi.org/10.1037/h0093718

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. http://doi.org/10.1126/science.1185718

Beckers, R., Deneubourg, J. L., & Goss, S. (1992). Trails and U-turns in the selection of a path by the ant Lasius niger. *Journal of Theoretical Biology*, *159*(4), 397–415. http://doi.org/10.1016/S0022-5193(05)80686-1

Boehm, C. (1996). Emergency decisions, cultural-selection mechanics, and group selection. *Current Anthropology*, *37*(5), 763–793. http://doi.org/10.1086/204561

Brainard, D. H. (1997). The Psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. http://doi.org/10.1163/156856897X00357

Brennan, A. A., & Enns, J. T. (2014). When two heads are better than one: Interactive versus independent benefits of collaborative cognition. *Psychonomic Bulletin & Review*, 1–7. http://doi.org/10.3758/s13423-014-0765-4

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, *214*(4516), 93–94.

Condorcet, M. J. A. N. de C. marquis de. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale.

Conradt, L., & Roper, T. J. (2003). Group decision-making in animals. *Nature*, *421*(6919), 155–158. http://doi.org/10.1038/nature01294

Conradt, L., & Roper, T. J. (2005). Consensus decision making in animals. *Trends in Ecology & Evolution*, *20*(8), 449–456. http://doi.org/10.1016/j.tree.2005.05.008

Couzin, I. D., Krause, J., Franks, N. R., & Levin, S. A. (2005). Effective leadership and decision-making in animal groups on the move. *Nature*, *433*(7025), 513–516. http://doi.org/10.1038/nature03236

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, *80*(2), 97–125. http://doi.org/10.1037/h0033951

Denkiewicz, M., Rączaszek-Leonardi, J., Migdał, P., & Plewczynski, D. (2013). Information-sharing in three interacting minds solving a simple perceptual task. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2172–2176). Austin, TX: Cognitive Science Society.

Dodge, Y. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press.

Eckstein, M. P., Das, K., Pham, B. T., Peterson, M. F., Abbey, C. K., Sy, J. L., & Giesbrecht, B. (2012). Neural decoding of collective wisdom with multi-brain computing. *NeuroImage*, *59*(1), 94–108. http://doi.org/10.1016/j.neuroimage.2011.07.009

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ernst, M. O. (2010). Decisions made better. *Science*, *329*(5995), 1022–1023. http://doi.org/10.1126/science.1194920

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. http://doi.org/10.1038/415429a

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451. http://doi.org/10.1038/075450a0

Geisler, W. S. (2003). Ideal observer analysis. In J. S. Werner & L. M. Chalupa (Eds.), *The visual neurosciences* (pp. 825–837). MIT Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508. http://doi.org/10.1037/0033-295X.112.2.494

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237. http://doi.org/10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. http://doi.org/10.1016/j.tics.2014.06.009

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology*, *10*(4), 371–384. http://doi.org/10.1016/0022-1031(74)90033-X

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2012). Effective integration of serially presented stochastic cues. *Journal of Vision*, *12*(8). http://doi.org/10.1167/12.8.12

Kalven, H., & Zeisel, H. (1966). *The American jury*. Boston: Little, Brown.

Kameda, T., Tindale, R. S., & Davis, J. H. (2003). Cognitions, preferences, and social sharedness: Past, present, and future directions in group decision making. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 458–485). Cambridge University Press.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*(1), 623–655. http://doi.org/10.1146/annurev.psych.55.090902.142009

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362. http://doi.org/10.1126/science.1216549

Kravitz, D. A., & Martin, B. (1986). Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, *50*(5).

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6).

Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology*, *41*(1), 585–634. http://doi.org/10.1146/annurev.ps.41.020190.003101

Lewin, K. (1947). Frontiers in group dynamics II. Channels of group life; social planning and action research. *Human Relations*, *1*(2), 143–153. http://doi.org/10.1177/001872674700100201

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. http://doi.org/10.1073/pnas.1008636108

McGrath, J. E., & Kravitz, D. A. (1982). Group research. *Annual Review of Psychology*, *33*(1), 195–230. http://doi.org/10.1146/annurev.ps.33.020182.001211

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. http://doi.org/10.1163/156856897X00366

Peterson, W. W., Birdsall, T. G., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171–212. http://doi.org/10.1109/TIT.1954.1057460

Petty, R. E., Harkins, S. G., Williams, K. D., & Latané, B. (1977). The effects of group size on cognitive effort and evaluation. *Personality and Social Psychology Bulletin*, *3*(4), 579–582. http://doi.org/10.1177/014616727700300406

Pratt, S. C., Mallon, E. B., Sumpter, D. J., & Franks, N. R. (2002). Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant Leptothorax albipennis. *Behavioral Ecology and Sociobiology*, *52*(2), 117–127. http://doi.org/10.1007/s00265-002-0487-x

Prins, H. H. T. (1996). *Ecology and behaviour of the African buffalo*. London: Chapman & Hall.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236. http://doi.org/10.1037/0096-3445.135.2.207

Seeley, T. D., & Buhrman, S. C. (1999). Group decision making in swarms of honey bees. *Behavioral Ecology and Sociobiology*, *45*(1), 19–31. http://doi.org/10.1007/s002650050536

Seeley, T. D., Camazine, S., & Sneyd, J. (1991). Collective decision-making in honey bees: How colonies choose among nectar sources. *Behavioral Ecology and Sociobiology*, *28*(4), 277–290. http://doi.org/10.1007/BF00175101

Simons, A. M. (2004). Many wrongs: the advantage of group navigation. *Trends in Ecology & Evolution*, *19*(9), 453–455. http://doi.org/10.1016/j.tree.2004.07.001

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. http://doi.org/10.1037/a0015145

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes*, *60*(1), 1–13. http://doi.org/10.1006/obhd.1994.1072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203. http://doi.org/10.1037/0033-295X.108.1.183

Sorkin, R. D., Luan, S., & Itzkowitz, J. (2008). Group decision and deliberation: A distributed detection process. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 464–484). John Wiley & Sons.

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group performance depends on the majority rule. *Psychological Science*, *9*(6), 456–463. http://doi.org/10.1111/1467-9280.00085

Stewart, K. J., & Harcourt, A. H. (1994). Gorillas' vocalizations during rest periods: Signals of impending departure? *Behaviour*, *130*(1/2), 29–40.

Sueur, C., & Petit, O. (2008). Shared or unshared consensus decision in macaques? *Behavioural Processes*, *78*(1), 84–92. http://doi.org/10.1016/j.beproc.2008.01.004

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

Tindale, R. S. (1989). Group vs individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes*, *44*(3), 454–473. http://doi.org/10.1016/0749-5978(89)90019-8

Vul, E., & Pashler, H. (2008). Measuring the crowd within Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. http://doi.org/10.1111/j.1467-9280.2008.02136.x

Ward, A. J. W., Sumpter, D. J. T., Couzin, I. D., Hart, P. J. B., & Krause, J. (2008). Quorum decision-making facilitates information transfer in fish shoals. *Proceedings of the National Academy of Sciences*, *105*(19), 6948–6953. http://doi.org/10.1073/pnas.0710344105

Watson, A. B., Barlow, H. B., & Robson, J. G. (1983). What does the eye see best? *Nature*, *302*(5907), 419–422. http://doi.org/10.1038/302419a0

Appendix A

Group Decision Rules (full descriptions)

Participants responded using a 10-point confidence scale as shown in Figure 2. Let **R** be a matrix of all the individual confidence ratings of a group, where $\mathbf{R}_{ij}$ is the confidence rating on the i*th* trial (rows) of the j*th* group-member (columns). Consequently, $\mathbf{R}_{i\bullet}$ is a row vector containing all three group-members' individual ratings on the i*th* trial, and $\mathbf{R}_{\bullet j}$ is a column vector containing all of the j*th* group-member's ratings across all trials. Hence, we can select all of the j*th* group-member's ratings during signal trials separately from noise trials and vice-versa using the following notation: $\mathbf{R}_{SIGNAL\ j}$ or $\mathbf{R}_{S\ j}$ for signal trials, and $\mathbf{R}_{NOISE\ j}$ or $\mathbf{R}_{N\ j}$ for noise trials.

Each rule results in a group decision variable D_i for each trial, and then makes a yes/no decision using a group decision criterion D_C. The group decision outcome on each trial O_i is "no" when $D_i < D_C$, and "yes" when $D_i > D_C$. Formally this is expressed as follows:

$$O_i = \left\{ \begin{array}{l} 0, \text{ if } D_i < D_C \\ 1, \text{ if } D_i > D_C \end{array} \right\}. \qquad \text{(A1)}$$

Criterion. For some rules (i.e., averaging, weighted linear combination, and Bayesian models), the group decision criterion D_C is a free parameter that we fit empirically for each group, and separately for the equal strength condition and the mixture condition. For the analysis assessing the relative effectiveness of each rule (Figure 3), we found the value of D_C that maximizes the *proportion correct* of the rule in question for each group (i.e., maximize the proportion of trials that the rule's yes/no decision is correct). For the analysis evaluating which rule the groups were actually adopting (Figure 4), we found the value of D_C that maximizes the *choice probability* of the rule in question for each group (i.e., maximize the proportion of trials that the rule's yes/no decision is the same as the group's actual collective yes/no response, irrespective of whether or not that yes/no decision is correct).

1. Majority

The group decision variable for each trial D_i is determined using a step function that compares the three participants' individual ratings on each trial to the explicit decision criterion of the 10-point confidence scale ($C = 5.5$) as follows:

$$D_i = \sum_{j=1}^{3} step\left(\mathbf{R}_{ij}\right) \text{ ; where } step(x) = \left\{ \begin{array}{l} +1, \text{ if } x > C \\ -1, \text{ if } x < C \end{array} \right\} . \quad \text{(A2)}$$

The outcome of the majority rule O_i is then determined using Eq. A1 with the group decision criterion $D_C = 0$. Note that unlike all the other rules we consider, the majority rule does not have any free fitting parameters.

2. Majority with exceptions

Let $\max\left(\mathbf{R}_{i\bullet}\right)$ denote the highest of the three individual ratings on each trial. The outcome of the majority with exceptions rule O_i follows the outcome of the majority rule, denoted O_i^{MAJ}, except when $\max\left(\mathbf{R}_{i\bullet}\right) \geq k$, where k is a free fitting parameter (k = 7, 8, 9, or 10, depending on the group), in which case the outcome is "yes" irrespective of what O_i^{MAJ} is. The value of k was separately fit for each group so as to maximize the *choice probability* of this rule during the *mixture condition*. Formally, this rule is expressed as follows:

$$O_i = \left\{ \begin{array}{c} O_i^{MAJ}, \text{ if } \max\left(\mathbf{R}_{i\bullet}\right) < k \\ 1, \text{ if } \max\left(\mathbf{R}_{i\bullet}\right) \geq k \end{array} \right\} . \quad \text{(A3)}$$

Note that the respective outcomes of this rule and the majority rule are often in good agreement with one another (when $\max\left(\mathbf{R}_{i\bullet}\right) < k$ or when $O_i^{MAJ} = 1$). In the main text, we refer to the subset of trials that they are in disagreement with one another (i.e., when $\max\left(\mathbf{R}_{i\bullet}\right) \geq k$ and $O_i^{MAJ} = 0$) as *conflict trials*, because the outcome of the majority with exceptions rule ("yes") is in conflict with the outcome of the majority rule ("no").

3. Averaging

The group decision variable for each trial D_i is determined using the mean of the three participants' individual ratings on each trial as follows:

$$D_i = \bar{\mathbf{R}}_{i\bullet} = \tfrac{1}{3} \sum_{j=1}^{3} \mathbf{R}_{ij} . \quad \text{(A4)}$$

The outcome of the averaging rule O_i is then determined using Eq. A1 with the group decision criterion D_C being fit as a free parameter.

4. Weighted linear combination

This rule takes into account the covariance of the group-members' ratings and how well each group-member discriminates between signal and noise trials (Sorkin & Dai, 1994). Critically, this rule is implemented using only the ratings from the first half of the experiment for the equal strength condition analysis, and only the ratings from the second half of the experiment for the mixture condition analysis. So, to be clear, this process is repeated twice for each group using the respective ratings that were recorded in each condition.

Let $\Delta\mu_{ij}$ contain the difference between the j*th* group-member's mean rating during signal trials and his or her mean rating during noise trials. Importantly, this mean difference is recalculated trial-by-trial after removing the ratings of the current i*th* trial. Formally, the mean difference row vector $\Delta\mu_{i\bullet}$ is calculated as follows:

$$\Delta\mu_{ij} = \bar{\mathbf{R}}_{SIGNAL\ j} - \bar{\mathbf{R}}_{NOISE\ j}\ ,\ j=1,2,3. \tag{A5}$$

The covariance matrix, denoted $\mathbf{\Sigma}$, is also recalculated trial-by-trial after removing the ratings of the current i*th* trial. Formally it is calculated as follows:

$$\mathbf{\Sigma}_i = \mathbf{\Sigma}_{SIGNAL} + \mathbf{\Sigma}_{NOISE}\ , \tag{A6}$$

where the covariance of the group-members' ratings is calculated separately for signal trials than from noise trials as follows:

$$\mathbf{\Sigma}_{SIGNAL} = \begin{bmatrix} \mathrm{Var}(\mathbf{R}_{S1}) & \mathrm{Cov}(\mathbf{R}_{S1},\mathbf{R}_{S2}) & \mathrm{Cov}(\mathbf{R}_{S1},\mathbf{R}_{S3}) \\ \mathrm{Cov}(\mathbf{R}_{S2},\mathbf{R}_{S1}) & \mathrm{Var}(\mathbf{R}_{S2}) & \mathrm{Cov}(\mathbf{R}_{S2},\mathbf{R}_{S3}) \\ \mathrm{Cov}(\mathbf{R}_{S3},\mathbf{R}_{S1}) & \mathrm{Cov}(\mathbf{R}_{S3},\mathbf{R}_{S2}) & \mathrm{Var}(\mathbf{R}_{S3}) \end{bmatrix};$$

$$\mathbf{\Sigma}_{NOISE} = \begin{bmatrix} \mathrm{Var}(\mathbf{R}_{N1}) & \mathrm{Cov}(\mathbf{R}_{N1},\mathbf{R}_{N2}) & \mathrm{Cov}(\mathbf{R}_{N1},\mathbf{R}_{N3}) \\ \mathrm{Cov}(\mathbf{R}_{N2},\mathbf{R}_{N1}) & \mathrm{Var}(\mathbf{R}_{N2}) & \mathrm{Cov}(\mathbf{R}_{N2},\mathbf{R}_{N3}) \\ \mathrm{Cov}(\mathbf{R}_{N3},\mathbf{R}_{N1}) & \mathrm{Cov}(\mathbf{R}_{N3},\mathbf{R}_{N2}) & \mathrm{Var}(\mathbf{R}_{N3}) \end{bmatrix}. \tag{A7}$$

To determine the weight that is assigned to each group-member's rating on each trial, we multiply the inverse of the covariance matrix (which is called a precision matrix (Dodge, 2006)) times the transpose of the mean difference row vector to elicit a weights vector as follows (Fukunaga, 1990):

$$w_{i\bullet}^{T} = \sum_{i}^{-1} \Delta\mu_{i\bullet}^{T} \ . \tag{A8}$$

The subscript i indicates that the weights vector, covariance matrix, and mean difference row vector are calculated separately for each trial after removing the ratings of the current i*th* trial (leave-one-out procedure).

Finally, the group decision variable for each trial D_i is determined using a weighted linear combination of the three participants' individual ratings on each trial as follows:

$$D_i = \sum_{j=1}^{3} w_{ij}\mathbf{R}_{ij} \ . \tag{A9}$$

The outcome of the weighted linear combination rule O_i is then determined using Eq. A1 with the group decision criterion D_C being fit as a free parameter.

5 and 6. Optimal Bayesian

The optimal Bayesian model is afforded knowledge about the statistical distribution of information across group-members. Specifically, during the equal strength condition, the model knows that all three group-members receive signals of equal strength. Conversely, during the mixture condition, the model knows that when the signal is present one random group-member receives strong evidence while the other two members receive weak evidence (although the model does not know the random trial-by-trial assignment of signal strengths to group members).

Using this information, we can express the posterior probability that the current i*th* trial is a signal (or noise) trial using Bayes theorem as follows (Geisler, 2003; Green & Swets, 1966; Knill & Richards, 1996; Peterson et al., 1954):

$$\mathrm{P}\left[signal \,|\, \mathbf{R}_{i\bullet}\right] \approx \mathrm{P}\left[\mathbf{R}_{i\bullet} \,|\, signal\right]\mathrm{P}\left[signal\right] \ , \tag{A10}$$

where $\mathrm{P}\left[\mathbf{R}_{i\bullet} \,|\, signal\right]$ is the likelihood of jointly eliciting the ratings of the current trial $\mathbf{R}_{i\bullet}$ given that it is a signal (or noise) trial, and $\mathrm{P}\left[signal\right]$ is the prior probability of a signal (or noise) trial. Importantly, participants were explicitly informed that the signal would be present 50% of the time, and so we can drop the prior term $\mathrm{P}\left[signal\right]$ when computing likelihoods as it does not affect the likelihood ratio for the group decision variable D_i (cf. Eq. A14).

To compute likelihoods, the model takes into account the covariance of the group-members' ratings similar to the weighted linear combination rule. Critically, the model uses only

the ratings from the first half of the experiment for the equal strength condition analysis and only the ratings from the second half of the experiment for the mixture condition analysis, and it omits the ratings of the current i*th* trial (leave-one-out procedure) when computing the means and covariances, similar to the weighted linear combination rule.

5. Equal strength condition: *Optimal Bayesian linear model*

Let μ_{SIGNAL} and μ_{NOISE} be mean row vectors that contain the respective mean rating of each group-member during signal trials and during noise trials respectively, while omitting the ratings of the current i*th* trial. Thus, for equal strength condition trials, the likelihood that the ratings of the current trial $\mathbf{R}_{i\bullet}$ (remember that $\mathbf{R}_{i\bullet}$ is a row vector) were jointly elicited during a signal trial is given as follows (Fukunaga, 1990; Peterson et al., 1954):

$$\mathrm{P}\left[\mathbf{R}_{i\bullet} \mid signal\right] = \tfrac{1}{Q}\exp\left(-\tfrac{1}{2}(\mathbf{R}_{i\bullet} - \mu_{SIGNAL})\sum\nolimits_{i}^{-1}(\mathbf{R}_{i\bullet} - \mu_{SIGNAL})^{\mathrm{T}}\right), \tag{A11}$$

where $\mathbf{\Sigma}_{i}^{-1}$ is the inverse of the covariance matrix of the group-members' ratings, and Q is the normalization constant of the multivariate probability distribution, which is given as follows:

$$Q = \sqrt{\left|\Sigma_{i}\right|\left(2\pi\right)^{n}}, \tag{A12}$$

where $\left|\Sigma_{i}\right|$ is the determinant of the covariance matrix, and n = 3 (group size). Similarly, the likelihood that the ratings of the current i*th* trial were jointly elicited during a noise trial is given as follows:

$$\mathrm{P}\left[\mathbf{R}_{i\bullet} \mid noise\right] = \tfrac{1}{Q}\exp\left(-\tfrac{1}{2}(\mathbf{R}_{i\bullet} - \mu_{NOISE})\sum\nolimits_{i}^{-1}(\mathbf{R}_{i\bullet} - \mu_{NOISE})^{\mathrm{T}}\right). \tag{A13}$$

The group decision variable for each trial D_i is determined using the ratio of the two likelihoods as follows:

$$D_{i} = \frac{\mathrm{P}\left[\mathbf{R}_{i\bullet} \mid signal\right]}{\mathrm{P}\left[\mathbf{R}_{i\bullet} \mid noise\right]}. \tag{A14}$$

The outcome of the optimal Bayesian linear rule O_i is then determined using Eq. A1 with the group decision criterion D_C being fit as a free parameter.

Note that for the equal strength condition, the optimal Bayesian linear model can be reduced to be identical to the weighted linear combination rule, assuming equal covariances for signal trials and noise trials (Green & Swets, 1966; Sorkin & Dai, 1994).

6. Mixture condition: *Optimal Bayesian non-linear model*

Unlike for equal strength condition trials where we assume equal covariances for signal trials and noise trials, for mixture condition trials we compute the covariance of noise trials separately from the covariance of signal trials. Furthermore, the optimal Bayesian non-linear model for the mixture condition takes into account the knowledge that, if it is a signal trial, only the j*th* group-member received strong evidence while the other two members received weak evidence. But because the model does not know the random trial-by-trial assignment of signal strengths to group members, the model calculates likelihoods for each of the three possible mutually exclusive sets of signal-strength assignments to group members: (i) weak, weak, strong; (ii) weak, strong, weak; and (iii) strong, weak, weak. Consequently, the mean row vector and covariance matrix are recalculated three times on each trial, using only the ratings of the trials when the j*th* member received the strong signal, and omitting the ratings of the current i*th* trial.

Let $\mu_{\text{j-Strong}}$ be a mean row vector that contains the respective mean rating of each group-member during trials when the j*th* group-member received strong evidence, and let $\Sigma_{\text{j-Strong}}$ be the covariance matrix of the group-members' ratings during these same trials. Thus, the likelihood that the ratings of the current trial $\mathbf{R}_{\text{i}\bullet}$ (remember that $\mathbf{R}_{\text{i}\bullet}$ is a row vector) were jointly elicited during a signal trial is given by summing likelihoods across all three possible mutually exclusive sets of signal-strength assignments to group members as follows:

$$\mathrm{P}\left[\mathbf{R}_{\text{i}\bullet} \mid signal\right] = \tfrac{1}{3}\sum_{\text{j=1}}^{3} \tfrac{1}{Q}\exp\left(-\tfrac{1}{2}(\mathbf{R}_{\text{i}\bullet} - \mu_{\text{j-Strong}})\Sigma_{\text{j-Strong}}^{-1}(\mathbf{R}_{\text{i}\bullet} - \mu_{\text{j-Strong}})^{\mathrm{T}}\right), \tag{A15}$$

where Q is the normalization constant of the multivariate probability distribution (cf. Eq. A12).

Conversely, the likelihood that the ratings of the current trial were jointly elicited during a noise trial is computed using the mean row vector that contains the respective mean rating of each group-member during noise trials, denoted μ_{NOISE}, and the covariance matrix of the group-members' ratings during these same trials, denoted Σ_{NOISE}, as follows:

$$\mathrm{P}\left[\mathbf{R}_{\text{i}\bullet} \mid noise\right] = \tfrac{1}{Q}\exp\left(-\tfrac{1}{2}(\mathbf{R}_{\text{i}\bullet} - \mu_{NOISE})\Sigma_{NOISE}^{-1}(\mathbf{R}_{\text{i}\bullet} - \mu_{NOISE})^{\mathrm{T}}\right), \tag{A16}$$

where Q is the normalization constant of the multivariate probability distribution (cf. Eq. A12).

The group decision variable for each trial $\mathrm{D_i}$ is determined using Eq. A14, and the outcome of the optimal Bayesian non-linear rule $\mathrm{O_i}$ is then determined using Eq. A1 with the group decision criterion $\mathrm{D_C}$ being fit as a free parameter.

7. Bayesian with uncertainty about signal strengths (for the mixture condition)

The optimal non-linear model in the previous section knows that, if it is a signal trial, there are only three possible sets of signal-strength assignments to group members during the mixture condition (cf. Eq. A15). The *Bayesian-with-uncertainty* model in this section does not know that, when the signal is present, one random group-member receives strong evidence while the other two members receive weak evidence. Instead, this model has a lot of uncertainty about how strong the signal is for each and every group member when it is present. Hence, this model considers many different sets of signal-strength assignments to group members. We note that unlike the optimal Bayesian non-linear rule that is able to compute the true covariance of the members' ratings when the j*th* group-member received strong evidence for each of the three possible sets of signal-strength assignments that it considers (i.e., $\Sigma_{\text{j-Strong}}$), this model uses the covariance of the members' ratings during noise trials (i.e., Σ_{NOISE}) as a proxy for the expected covariance for each of the numerous sets of signal-strength assignments that it considers.

To introduce uncertainty about signal strengths, we force this model to consider many different mean row vectors when computing the likelihood that it was a signal trial, compared to the optimal non-linear rule that considers only three different mean row vectors (i.e., $\mu_{\text{j-Strong}}$) because it knows that there are only three possible sets of signal-strength assignments. Of course there are countless ways of generating many different mean row vectors to create uncertainty about signal strengths. Here we decided to use the group-members actual ratings during the mixture condition to create reasonable mean row vectors that "could have been plausible" given the ratings that group members actually exhibited during the mixture condition. Specifically, this model will consider X number of equally-spaced steps starting at each group-member's true mean rating when he/she received strong evidence during signal trials, down until (but not including) his/her mean rating during noise trials. These steps are calculated separately for each trial based on the member's ratings on all mixture condition trials except for the current trial (leave-one-out procedure). After calculating the respective steps for each group-member, the Bayesian-with-uncertainty model will then consider every single combination possible, which leads to X^3 = H number of mean row vectors, denoted μ_h. For the analysis that is shown in Figure 10, we chose to use 15 steps, thus generating 3375 different mean row vectors that the Bayesian-with-uncertainty model must consider ($15^3 = 3375$).

The likelihood that the ratings of the current trial $\mathbf{R}_{i\bullet}$ were jointly elicited during a signal trial is given by summing likelihoods across all H mutually exclusive mean row vectors as follows:

$$P\left[\mathbf{R}_{i\bullet} \mid signal\right] = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{Q}\exp\left(-\frac{1}{2}(\mathbf{R}_{i\bullet}-\mu_h)\sum\nolimits_{NOISE}^{-1}(\mathbf{R}_{i\bullet}-\mu_h)^T\right), \quad \text{(A17)}$$

where Q is the normalization constant of the multivariate probability distribution (cf. Eq. A12). The likelihood that the ratings of the current trial were jointly elicited during a noise trial is computed using Eq. A16. The group decision variable for each trial D_i is determined using Eq. A14, and the outcome of the Bayesian-with-uncertainty rule O_i is then determined using Eq. A1 with the group decision criterion D_C being fit as a free parameter.

8. Associative heuristic (for the mixture condition)

The optimal Bayesian model knows that when the signal is present during the mixture condition, one random participant receives strong evidence while the other two participants receive weak evidence. This prompts the optimal Bayesian non-linear rule to regularly go against the majority opinion (and occasionally even against a unanimous opinion) and respond "no" when no-one in the majority is very confident that it was a signal trial. A simple way to approximate this feature is to refrain from responding "yes" *unless* someone is very confident that it was a signal trial during the mixture condition. This heuristic could potentially be learned associatively by noticing, through the feedback, that when the signal is present someone is usually very confident that it was a signal trial. Note that this associative heuristic is similar to the majority with exceptions rule; but unlike the majority with exceptions rule that only bucks the majority rule to respond "yes" when someone is very confident that is was a signal trial, this associative heuristic rule will also buck the majority rule to respond "no" when *no-one* is very confident that it was a signal trial. Indeed, this associative rule will even go against a unanimous opinion if no-one is very confident that it was a signal trial. Critically, this rule goes beyond the majority with exceptions rule to completely ignore what the majority has to say and to only pay attention whether or not someone is very confident that the signal was present.

Let $\max\left(\mathbf{R}_{i\bullet}\right)$ denote the highest of the three individual ratings on each trial. The outcome of this rule is "no" except when $\max\left(\mathbf{R}_{i\bullet}\right) \geq k$, where k is a free fitting parameter (k = 7,

8, 9, or 10, depending on the group), in which case the outcome is "yes". The value of *k* was separately fit for each group so as to maximize the *choice probability* of this rule during the *mixture condition*. Note that the value of *k* for each group is not necessarily the same for this rule as in the majority with exceptions rule. Formally, this rule is expressed as follows:

$$O_i = \left\{ \begin{array}{l} 0, \text{ if } \max\left(\mathbf{R}_{i\bullet}\right) < k \\ 1, \text{ if } \max\left(\mathbf{R}_{i\bullet}\right) \geq k \end{array} \right\}. \quad \text{(A18)}$$

Note that unlike the majority with exceptions rule, this rule does not pay attention at all to the majority opinion because all it cares about is the value of $\max\left(\mathbf{R}_{i\bullet}\right)$; i.e., the value of the highest individual rating.

Appendix B

Details of the Theoretical Simulations in Figure 3

The simulations were for groups of three members and we compared the outcomes of different group decision algorithms in different environments. For each simulation, we generated 300,000 noise trials and 300,000 signal trials. For each trial we drew three correlated random variables (one for each member), with the correlation set to .2. The random variables were all drawn from Gaussian distributions with unit standard deviation. The underlying Gaussian for noise trials throughout the simulations was always centered on zero ($\mu_{NOISE} = 0$). We controlled the signal to noise ratio (SNR), which equals d' (cf. Eq. 3 in the main text), by shifting the underlying Gaussian for signal trials away from zero (SNR = μ_{SIGNAL}).

For the equal strength situation simulations, the underlying Gaussian was shifted to the same location for all three members. The simulations show the outcomes of different integration algorithms as a function of the location of μ_{SIGNAL}, ranging from 0.5 to 3 in 0.1 intervals.

For the mixture situation simulations, each signal trial consisted of two members receiving random variables that are drawn from a Gaussian that is located at $\mu_{WEAK} = 0.5$, and one member receiving a random variable from a Gaussian that is located at μ_{STRONG}. Each member receives μ_{STRONG} on a different third of the signal trials, which means that each member

gets μ_{WEAK} on 200,000 trials and μ_{STRONG} on 100,000 trials. The simulations show the outcomes of different integration algorithms as a function of the location of μ_{STRONG}, ranging from 0.5 to 3 in 0.1 intervals. Note that when $\mu_{STRONG} = \mu_{WEAK} = 0.5$, the simulated results are the same as in the simulation of the equal strengths situation because all three members are getting signals of equal strength with SNR = 0.5.

The decision criterions in the simulations were set so as to maximize *proportion correct*. The "Majority" and "Majority with exceptions" rules require setting decision criterions at the individual-response level so as to maximize each individual's total number of correct individual responses. For these simulations, we chose to invoke the exception for the "Majority with exceptions" rule when any individual's random variable is more than 1.5 standard deviations greater than their optimally set criterion (this is akin to being very confident that the signal is present, and so this rule responds "yes" even if the other two individuals' random variables are lower than their respective criterions). The "Averaging" and "Weighted linear" rules require setting decision criterions at the group response level so as to maximize the total number of correct group responses. For the "Optimal" rule, we use the theoretically optimal decision criterion.