Supplementary Material

In Study 1, to cast doubt on the possibility that perceivers attended to a single face to drive their judgments rather than genuinely extracted the ensemble mean, we simulated participants attending to a single face at random from the ensemble. For every trial of every participant, we randomly selected a face from the ensemble and calculated the absolute distance in trustworthiness between the randomly selected face (random distance) and the ensemble mean (mean distance). We then conducted a mixed-effects model predicting the likelihood of a correct response from random distance and mean distance with random intercepts for participant and ensemble stimuli (formula: correct ~ 1 + *absolute ensemble trustworthiness difference* + *absolute random face trustworthiness difference*). We extracted the coefficients (log-odds) for the two predictors and repeated this process 1,000 times resulting in 1,000 coefficients for each predictor. A paired t-test between the two coefficients indicated that performance increased as a function of distance between the ensemble mean and the probe *(M*=0.358, *SD*=0.006), but not as a function of the distance between a single random face and the probe *(M*=-0.010, *SD*=0.011); *t*(999)=676.57, *SE*= 95% CI [0.367, 0.369], p<0.00001.

As a complementary analysis, we simulated null distributions for the distance effect in individual subjects. The previous results show that correct trials are associated with greater distance between the ensemble mean and the probe than incorrect trials. If participants are engaging in ensemble coding, this [incorrect – correct] distance effect should be greater than when calculated using random distances. Splitting correct and incorrect trials into two pools for each subject, we sampled 100 random distances each from the incorrect and correct pools by selecting one of the 8 faces on every trial. We then averaged these 100 values for each pool and calculated the [incorrect – correct] effect, thereby making it commensurate with the true effect (as the true effect is an average of 100 trials). We repeated this process 1,000 times to generate a null distribution for each subject and tested whether each subject's true effect exceeded 95% of the null distribution. A greater proportion of subjects showed a significant effect on an individual subject basis (50.2%) than would be expected by chance (5%) (exact binomial test, p<0.00001, two-tailed). In addition, across subjects, the true [incorrect – correct] grand mean distance effect also exceeded 91% of the aggregated null distribution, a finding that is highly unlikely to occur by chance. Together, these two analyses strongly suggest that participants extracted the ensemble mean rather than attended to a face at random to infer the ensemble's trustworthiness.