COLLECTIVE INTELLIGENCE

Research article

Collective Intelligence
Vol. 5(1): 1–10

DOI: 10.1177/26339137261435121
journals.sagepub.com/home/col
Sage

Three and a half stages of crowd wisdom

Igor Douven[1], Nikolaus Kriegeskorte[2] and Patrick Stinson[2]

Abstract

Background: According to the wisdom of the crowd idea, crowds can be smart even when most of their members are not. Much of the literature assumes that wisdom is to be extracted from a crowd by averaging the numerical predictions or probabilistic estimates of its members.
Purpose: We compare a broad range of aggregation procedures, from simple averaging methods to optimized dynamic models and neural networks.
Research Design: We apply multiple aggregation methods — including simple averaging, optimized weighted averages, a dynamic Hegselmann–Krause model of social learning, and two neural network architectures — to a dataset of probabilistic judgments, evaluating performance by both classification accuracy and Brier scores.
Study Sample: 376 individuals providing probabilistic judgments about 1,200 statements with known ground truth.
Data Collection and/or Analysis: Performance of all aggregation methods was assessed using classification accuracy and Brier scores, with neural network architectures trained to learn an optimal aggregation function directly from the data.
Results: More sophisticated aggregators yield better performance. While some simple averaging methods exceed the performance of the average participant, optimized weighted averages and the Hegselmann–Krause model achieve significantly higher performance. Two neural network architectures outperform all other methods by a large margin, reaching a level of accuracy vastly superior to that of even the best individual in the crowd.
Conclusions: Crowd wisdom is best not thought of as a fixed property but rather as something that can be achieved to different degrees, depending on the method used for aggregating opinions.

Keywords

agent-based models, artificial neural networks, collective intelligence, Hegselmann–Krause model, probability aggregation, wisdom of crowds

Introduction

The wisdom of the crowd idea is the claim that crowds can be smart even when their members are not (Becker et al., 2019). This broad idea has been made precise in two different ways: that the crowd's aggregate opinion is better than that of its "average member" (see, e.g., Galton, 1907; many examples in Surowiecki, 2004; Davis-Stober et al., 2014), and that the aggregate opinion is better than that of even the crowd's best (most expert or accurate) members (Hong and Page, 2004, 2025; Page, 2007; Prelec et al., 2017). While there is some evidence for both claims (see, e.g., the books by Surowiecki and Page), there is also evidence *against* the wisdom of the crowd idea generally (in that there is evidence even against the weaker, better-than-average reading). It has been found, for instance, that social influence can lead to herding and thus reduce collective accuracy (Lorenz et al., 2011; but cf. Navajas et al., 2018).

The mixed evidence has raised interest in the conditions under which crowds can be expected to be wise (e.g., the degree to which the opinions of members can be correlated

[1]SND / CNRS / Sorbonne University, Paris, France
[2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

Corresponding author:
Igor Douven, SND / CNRS / Sorbonne University, 1 rue Victor Cousin, Paris 75005, France.
Email: igor.douven@sorbonne-universite.fr

or the presence of different kinds of bias; see Davis-Stober et al., 2014). Here, we are interested in a different question, having to do with the aggregation method that is used to *extract* wisdom from a crowd. Much research on crowd wisdom only considers the simple average (i.e., the arithmetic mean) as an aggregation method. In this paper, we go beyond that by considering a range of increasingly more sophisticated aggregation methods and testing them empirically, using a large dataset of probabilistic judgments from 376 participants on 1200 statements with known ground truth. Our results show that we can achieve different levels of crowd wisdom, some even yielding estimates that top those, not just of the "average," but of the best participant.

In a preliminary stage, we conduct a baseline analysis by considering the simplest possible generalization of the standard approach to crowd wisdom, also looking at the geometric and harmonic mean. In the first real stage, we move to weighted averages, finding optimal weights for each participant. The main contributions of this paper are to be found in the second and third stages. In the second, we consider iterated averaging via a known model of opinion dynamics, specifically, the Hegselmann–Krause model (Hegselmann and Krause, 2002). And in the third stage, we go beyond predefined formulas and train two types of neural network to learn an optimal aggregation function directly from the data. We find that most methods do better than the average participant but that only the neural network aggregators outperform *every* individual participant.

We start by detailing the hierarchy of opinion aggregators we compare (Sect. 2), before presenting our empirical results (Sect. 3). As we will show, the method of aggregation is a decisive factor in determining the stage of wisdom a crowd can achieve. Finally, we discuss the broader implications of these findings, also pointing out some limitations of our work and mentioning some avenues for further research (Sect. 4).

Aggregators

"Wisdom of the crowd" is the heading for a broad thesis that is generally taken to apply to very different types of judgments. Classic examples include aggregate estimates of continuous quantities like the weight of an ox (Galton, 1907) or the height of the Eiffel tower (Navajas et al., 2018), but also aggregate categorical predictions—for instance, concerning which candidate will win an election—and rankings, such as ordering items by preference or quality (e.g., Landemore, 2012). Our data consist of probability judgments, and so we will be specifically interested in whether aggregating probability judgments—using various aggregation methods—yields "better than average" or even "better than best" probabilities. Naturally, this has consequences for which aggregators we consider; for instance, we do not consider voting rules, which operate on categorical judgments. Here are the aggregators to be considered in our study:

Simple averages

What is commonly referred to as "the average" is only one member of a family of means, sometimes called the *f-means* and defined as follows (Bullen, 2003, p. 266):

$$M_f(x_1, \ldots, x_n) =_{\mathrm{df}} f^{-1}\left(\frac{\Sigma_{i=1}^{n} f(x_i)}{n}\right)$$

If we let f be any linear function (e.g., the identity function), this yields the arithmetic mean; if we let f be the logarithmic function, we obtain the geometric mean; and if we let f be the reciprocal function, we get the harmonic mean. These are arguably the best known means, and we consider them all. In addition, we consider the log-odds mean, which is the instance of the above schema with f the logit function (and so f^{-1} the logistic function), given that it has been advocated as the normative standard for probabilistic aggregation under the assumptions that individual judgments are conditionally independent and well-calibrated (see Genest and Zidek, 1986; Morris, 1983;).

Weighted averages

A still broader family of means are the *weighted f*-means, defined as

$$M_f^{\omega}(x_1, \ldots, x_n) =_{\mathrm{df}} f^{-1}\left(\frac{\Sigma_{i=1}^{n} \omega_i f(x_i)}{\Sigma_{i=1}^{n} \omega_i}\right)$$

where $\omega = (\omega_1, \ldots, \omega_n)$ are nonnegative weights (Bullen, *ibid.*). While equal-weight averaging (which is what the *f*-means do) treats all participants equally, one may want to weight participants based on their overall previous performance or on some other property. In Section 3, we proceed differently by recruiting an evolutionary algorithm to find participant weights that maximize performance. There, we consider weighted versions of the four means mentioned previously.

Dynamic aggregation

The Hegselmann–Krause model. Moving beyond static aggregation methods, we implement a modified version of the Hegselmann–Krause (HK) model of opinion dynamics (Hegselmann and Krause, 2002). Originally developed to study consensus formation and polarization in social networks, the HK model treats opinion aggregation as an iterative process in which agents update their opinions based on those of others that are "similar enough" to them (their "peers").[1] More specifically, the model considers a community V of agents whose opinions can all be represented by a real number in the unit interval. At each iteration, an agent

updates its opinion by averaging those of its peers at the given iteration, where peerhood is formalized through the notion of a *bounded confidence interval* (BCI). An agent i's BCI encompasses those agents whose opinions differ from i's own opinion by no more than some value $\epsilon_i \in [0, 1]$, referred to as agent i's *confidence bound*. Formally, at each iteration s, an agent i updates its opinion according to

$$x_i(s+1) = \frac{1}{|X_i(s)|} \sum_{j \in X_i(s)} x_j(s),$$

with $x_j(s)$ the opinion of agent j after the update at iteration s and

$$X_i(s) =_{\text{df}} \{j \in V : |x_i(s) - x_j(s)| \leqslant \epsilon_i\}$$

the set of agents within agent i's BCI after the update at s.

Note that in the standard HK model, every agent is always one of its own peers and also that, when updating, its opinion has exactly the same weight as those of its peers. As Fu et al. (2015) argue, that is not always realistic. In reality, we will often want to attach more weight to our own opinion, even if we also want to take into account those of others. They therefore propose a modification of the HK model in which each agent i is characterized by two parameters: $\alpha_i \in [0, 1]$ and $\epsilon_i \in [0, 1]$, where the former regulates the self-weight in opinion updating (i.e., how much weight the agent gives to its own opinion relative to the weight assigned to the average of its peers' opinions) and the latter is again the threshold for peer selection based on opinion distance, although now the agent itself is excluded from counting as a peer. Their model then iteratively updates opinions according to

$$x_i(s+1) = \begin{cases} \alpha_i x_i(s) + (1-\alpha_i)\frac{1}{|\breve{X}_i(s)|}\sum_{j \in \breve{X}_i(s)} x_j(s), & \text{if } \breve{X}_i(s) \neq \varnothing \\ x_i(s), & \text{otherwise} \end{cases}$$

with

$$\breve{X}_i(s) =_{\text{df}} \{j \in V \setminus \{i\} \mid |x_i(s) - x_j(s)| \leqslant \epsilon_i\}.$$

As presented here, both (HK) and (MHK) are *one-dimensional* models. However, opinions in the real world are rarely isolated. As a result, an individual's belief state is best represented as a vector in a multi-dimensional opinion space. The Hegselmann–Krause framework was explicitly meant to be extensible to such multi-dimensional cases (Hegselmann and Krause, 2002), and various multi-dimensional extensions have been proposed and analyzed (Amblard and Deffuant, 2004; Chen et al., 2019; Deffuant et al., 2000; Douven and Hegselmann, 2022; Fortunato et al., 2005; Huet et al., 2008; Lorenz, 2007, 2008). These models generalize by replacing the one-dimensional opinion distance $|x_i - x_j|$ with a multi-dimensional distance metric $d(x_i, x_j)$ over the opinion vectors. We adopt this approach, applying it to the multi-dimensional generalization of (MHK). Specifically, we define the distance between two agents as the mean squared difference between their probability judgments across all 1200 claims:

$$d_{ij} =_{\text{df}} \frac{1}{1200} \sum_{k=1}^{1200} \big(p_i(k) - p_j(k)\big)^2.$$

This distance matrix, which reflects systematic patterns of agreement rather than claim-specific proximity, determines peerhood: agent j is a peer of agent i if and only if $d_{ij} \leqslant \epsilon_i$. Importantly, we recompute this distance matrix at each iteration based on agents' current (updated) probabilities, so that peer relationships can evolve as opinions converge.

Finally, while the first clause in Fu and colleagues' model takes a weighted sum of the agent's own opinion and the *arithmetic* mean of its peers' opinions, one could use other means here. In our study, we consider again, next to the arithmetic mean, the geometric and harmonic means,[2] as well as the log-odds mean. For each of these four averaging variants, we optimize the $2N$ parameters (α_i and ϵ_i for each of $N = 376$ participants) on training data using multi-objective evolutionary optimization targeting both Brier score minimization and accuracy maximization (see below).

Neural network aggregators. Finally, we go beyond predefined aggregation formulas to allow machine learning models, specifically artificial neural networks, to learn optimal aggregation functions directly from the data. We implement two architectures:

- Multilayer perceptron (MLP): The MLP we will use consists of three fully connected layers with sizes $376 \to 64 \to 32 \to 1$, where 376 is the input dimension (corresponding to the number of participants), and the two hidden layers have 64 and 32 units, respectively. We use ReLU activation functions for the hidden layers and a sigmoid activation function for the output layer to ensure probability outputs. To avoid overfitting, we incorporate dropout regularization with a rate of 0.2 after each hidden layer. The network will be trained using the Adam optimizer with a learning rate of 0.001, minimizing binary cross-entropy loss over 25 epochs with a batch size of 16.
- Long short-term memory (LSTM) network: In our study, the recurrent architecture consists of two stacked LSTM layers with hidden sizes 32 and 16, respectively, taking as input sequences of vectors of dimension 376 (corresponding to the number of participants). The final LSTM layer is followed by a dense output layer with sigmoid activation. We will process each input through the network for 5 time steps, allowing the model to iteratively refine its predictions. This is meant to reflect somewhat the

iterative updating that occurs in the MHK model. We use the RMSProp optimizer with default parameters, training for 25 epochs with the same batch size and loss function as the MLP.

We selected these relatively simple architectures after preliminary experiments showed that larger networks did not provide additional benefit on this task, probably due to the limited training data (960 claims, being 80 % of the 1200 claims on which the dataset was based; see below). Both neural networks were implemented in Julia using the `Flux.jl` deep learning framework. For further computational details, readers are referred to the Supplemental Materials.

Study

Materials and procedure

To test the various aggregation methods, we reanalyzed data from Stinson et al. (2025), which consist of probabilities assigned by 376 participants to 1200 general knowledge claims with known ground truth (e.g., "The first handheld calculator was built before 1975," "Mexico was the main trading partner of the US in the 1990s"). The claims fall into six categories (history, geography, science, sports and leisure, social sciences and politics, arts and entertainment), with 200 claims per category, 100 of which are true and 100 false. Participants evaluated the claims in six online sessions, in each of which 200 unique claims were presented, in an order randomized per participant. Participants were not informed of the base rate.[3]

As mentioned, that our data consist of probability judgments has consequences for which aggregation methods we can consider. At the same time, it has consequences for how to measure performance and thus how to determine whether one aggregator is better than another. For this purpose, we used two well-known criteria: (i) accuracy (or classification correctness), which is the proportion of binarized probabilities (true if $>.5$, false otherwise) that coincide with the actual truth values (which we know, in the case of our data); and (ii) the Brier score, which is the squared difference between the probability assigned to a claim and that claim's actual truth value (0 for false, 1 for true). Note that for the former criterion, higher is better, while for the latter, lower is better.

To ensure that our results are generalizable, we divided the 1200 claims into a training set (80 %) and a held-out test set (20 %). The final performance of all aggregators is reported on the test set.

For the models requiring optimization—the weighted averages and the MHK models—parameters were fit exclusively on the training set. To do this, we used a multi-objective evolutionary algorithm, specifically the Borg MOEA (Hadka and Reed, 2013).[4] In our case, the objectives for the algorithm to simultaneously optimize were minimizing the aggregate Brier score and maximizing the aggregate classification accuracy. The algorithm evolves a population of "candidate solutions" (i.e., vectors of parameters) over many generations to discover a set of non-dominated, or Pareto-optimal, solutions. However, it turned out that our objectives were sufficiently aligned that optimization consistently converged to a single Pareto-optimal solution rather than a frontier of trade-offs.[5]

To assess the stability and uncertainty of our models' performance, we employed validation procedures appropriate to each model's structure and computational cost. For the simple and weighted averaging methods, as well as the MHK models, we conducted a bootstrap analysis by repeatedly resampling the 376 participants with replacement from the test set ($N = 1,000$ resamples for simple/weighted averages; $N = 500$ for the MHK models) and recalculating the performance metrics. For the neural network models, we assessed stability by training each architecture 100 times from different random initializations and we report the mean and standard deviation of the performance on the test set. This mixed-method approach to validation provides robust uncertainty estimates while remaining computationally feasible, given that a full bootstrap re-optimization was, at least for the more complex models, computationally much too costly.[6]

For the optimization of the parameters of the MHK model, we simulated the model for 100 iterations within each call of the objective function. As can be seen in the Supplemental Materials, analysis of the full opinion dynamics over 250 iterations revealed that both the Brier score and accuracy of the aggregate opinion largely converge within the first 100 iterations. Extending the simulation during the optimization phase would have substantially increased the computation time with little expected improvement in the quality of the discovered parameters. In the bootstrap procedure for this model, we chose the same number of iterations.

Results

The performance of all 11 aggregation procedures on the held-out test data is summarized in Figure 1. The results reveal a clear trend: as the sophistication of the aggregation method increases, collective performance, measured by both mean Brier score (lower is better) and classification accuracy (higher is better), improves.

Pre-stage: Simple averages. We first established a baseline using simple, unweighted *f*-means (blue bars in Figure 1). In this category, the geometric mean (GM) emerged as the winner, achieving a mean Brier score of 0.143 (95 % CI [0.139, 0.148]) and an accuracy of 0.829 (95 % CI [0.806, 0.853]). The other three means performed significantly worse, on both criteria (as confirmed by a series of *t*-tests,

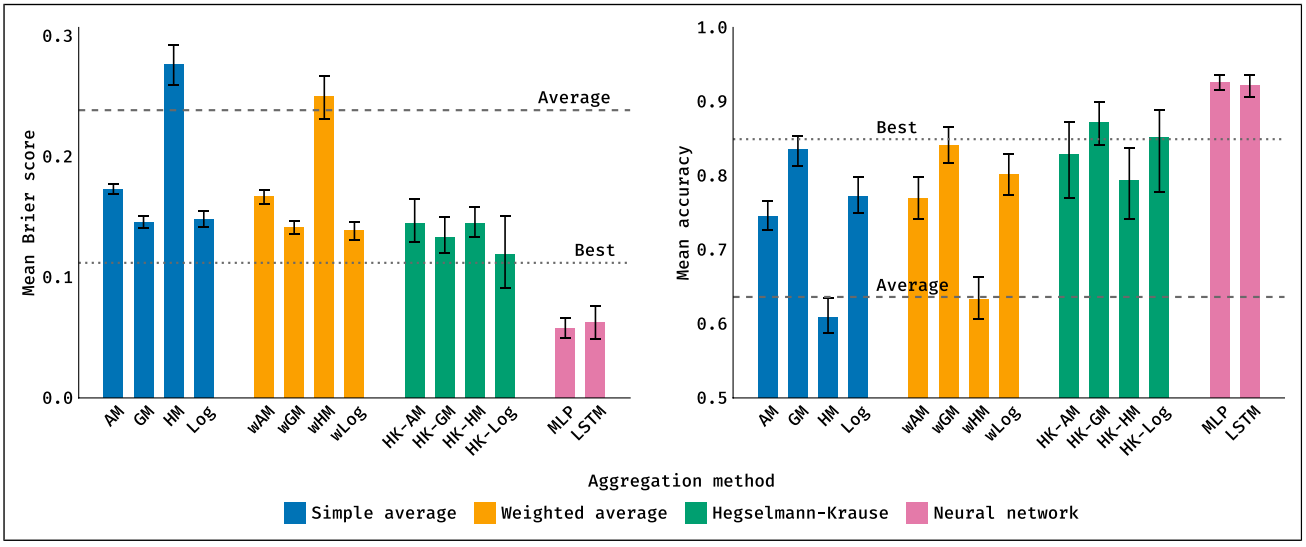

Figure 1. Mean Brier score (left panel, lower is better) and mean classification accuracy (right panel, higher is better) for 11 aggregation methods on the held-out test set. Methods are grouped by type: simple unweighted averages (blue), optimized weighted averages (orange), MHK dynamic models (green), and neural networks (pink). Error bars represent 95 % confidence intervals from bootstrap analysis (for the first three groups) or ±1 standard deviation from 100 independent training runs (for the neural networks). Horizontal lines indicate the performance of the average individual participant (dashed) and the best individual participant (dotted) in the test set (where the best individual was different for the two criteria). The results show a clear trend of improving performance with increasing aggregator sophistication, culminating in the neural networks, which are the only methods to do better than the best individual on both criteria.

with all ps < .0001 and all associated effect sizes, measured via Cohen's d, being large). While the log-odds mean is often considered a normative benchmark (Morris, 1983), that it does not come out on top here simply suggests that the model's core assumption—that individual judgments are well-calibrated and conditionally independent—are violated in our data set, which is a common issue with real-world human data (Genest and Zidek, 1986). Nevertheless, the arithmetic, geometric, and log-odds means all support the "better-than-average" claim of crowd wisdom, as their performance exceeded that of the average individual participant (dashed line). However, none of these simple aggregators approached the performance of the crowd's best individual member (dotted line).[7]

Stage I: Weighted averages. In the first stage of sophistication, we used the Borg MOEA evolutionary algorithm described above to find an optimal set of weights for each participant, moving from equal to differential weighting (orange bars). An optimization procedure was run for each of the arithmetic, geometric, harmonic, and log-odds means. The optimally weighted geometric (wGM) and log-odds (wLog) means were ex-aequo winners, the former doing significantly better than the latter with respect to accuracy (0.841, with 95 % CI [0.818, 0.865], vs 0.8, with 95 % CI [0.774, 0.829]; t < .0001, d = 5.35), while for the other criterion it was the other way around (0.14, with 95 % CI [0.131, 0.146] vs 0.141, with 95 % CI [0.136, 0.147]; t < .0001, d = 0.83). Both weighted means achieved a significantly better mean Brier score than the simple geometric mean, as confirmed by two t-tests (p < .0001 for both, with Cohen's d equal to 1.75 for wGM and equal to 2.31 for wLog), and the weighted geometric mean (but not wLog) also achieved a significantly higher accuracy than the simple geometric mean (t < .0001, d = 0.59). The other two optimally weighted averages did significantly worse than the weighted geometric and log-odds means but significantly better than their unweighted counterparts. Overall, the results confirm that even a simple form of data-driven optimization—learning to give more influence to some participants—can extract additional wisdom from the crowd.[8] Interestingly, correlations between, on the one hand, optimal weights for any of the means and, on the other, either participants' accuracy or their average Brier score was invariably low, with all $|r|$ < .32.

Stage II: Dynamic aggregation. In the second stage, we moved from static aggregation to a dynamic, iterative process modeled on social learning, using the modified Hegselmann–Krause (MHK) model introduced earlier (green bars). As mentioned, we used the Borg MOEA algorithm to find optimal α and ϵ values for the participants, running a separate optimization procedure for each of the means of interest.

The shift to iterative averaging produced a significant leap in performance. The MHK model that used the

log-odds mean to average peer opinions (HK-Log) achieved the lowest Brier score in this group at 0.119 (95 % CI [0.0912, 0.151]), an improvement that was statistically significant compared to the (on this criterion) best weighted average model wLog ($p < .0001$, $d = 2.74$). On the other hand, the MHK model using the geometric mean did best on the other criterion, achieving an accuracy of 0.872 (95 % CI [0.841, 0.899]), which was a significant improvement over the (on this criterion) best weighted average model wGM ($p < .0001$, $d = 3.1$). These results indicate that modeling a structured social interaction, where agents update their beliefs based on a network of trusted peers, is a more powerful mechanism for aggregation than static weighting alone. The underlying opinion dynamics of this process is illustrated in the Supplemental Materials.

Similar to the correlation analysis in the first stage, we checked whether there was a relation between the optimal α and ϵ parameters for the given means and the participants' accuracies and average Brier scores running a linear regression with either α parameters or ϵ parameters as dependent variable, and with accuracies and average Brier scores as predictors, also including ϵ parameters (when α parameters were the dependent variable) or α parameters (when ϵ parameters were the dependent variable) as covariate. In none of these regressions did any of the predictors turn out to be significant.

It is further noteworthy that the optimized MHK parameters revealed substantial heterogeneity across participants. The Supplemental Materials contains plots of the various distributions, showing that both α and ϵ parameters span nearly the full [0, 1] range for all pooling variants, with no strong clustering patterns. This indicates that some participants optimally retain high self-weight while others defer substantially to peers, and that some benefit from broad confidence bounds (interacting with diverse others) while others perform best with selective peer groups.

Stage III: Neural network aggregators. In the final stage, we abandoned predefined formulas entirely and trained two neural network architectures to learn an optimal aggregation function directly from the data (pink bars). This approach yielded by far the best performance. The multilayer perceptron (MLP) and the long short-term memory (LSTM) network achieved nearly identical results. The MLP reached a mean Brier score of 0.068 (±0.008) and a mean accuracy of 0.913 (±0.011), while the LSTM network reached a mean Brier score of 0.076 (±0.018) and a mean accuracy of 0.906 (±0.021). On both criteria, the MLP scored significantly better (both ps $< .005$), but the effect size was small in both cases (both ds < 0.46). However, both network architectures significantly improved, given either of our criteria, over the most accurate MHK model HK-GM (both ps $< .0001$, $d = 2.23$ for the MLP, $d = 1.58$ for the LSTM network) as well as over HK-Log, which is the MHK model with the lowest Brier scores (both ps $< .0001$, $d = 3.95$ for the MLP, $d = 3.51$ for the LSTM network).

It is important to note that the neural network approach is the only one that decisively surpasses the performance of the crowd member with best Brier score and the crowd member with best accuracy, showing that a non-linear aggregator (and perhaps only such aggregators) can reach the "better-than-best" level of crowd wisdom. The similar performance of the two distinct architectures also suggests that this finding is robust and not an artifact of a specific model choice.[9]

Discussion

Our study systematically investigated how collective performance depends on the choice of aggregation method. The results present a clear picture: the wisdom that can be extracted from a crowd is not a fixed quantity, but is contingent on the sophistication of the aggregator. Moving from simple averages to optimized weights, then to dynamic social models, and finally to learned functions, we observed a statistically significant improvement in performance at each stage (Figure 1).

The most significant finding is the performance of the neural network aggregators, which provides strong empirical support for the "better-than-best" conception of crowd wisdom (Hong and Page, 2004; Page, 2007). While simpler methods confirmed that crowds can outperform their average member, only a non-linear aggregator was able to synthesize the judgments of 376 individuals into a final prediction that was superior to that of the single best crowd member. This suggests that in settings like ours, where the available inputs are limited to forecasters' raw probability judgments, the highest levels of collective intelligence may be achievable most readily when (and perhaps even only when) we go beyond predefined formulas and allow an aggregation mechanism to learn complex, non-obvious patterns of reliability and interaction directly from the data. The robustness of this finding, demonstrated by the nearly identical performance of two different neural architectures, further strengthens this conclusion.

The same finding also establishes a conceptual and empirical bridge between the study of collective intelligence in social science and the principles of ensemble learning in machine learning.[10] In that area, it has long been known that combining multiple, diverse models often yields a better performance than any individual model on its own (Breiman, 1996; see also Huang et al., 2024; Schoenegger et al., 2024). Our results show that this principle also holds when the "individual models" are human agents and we conceptualize a crowd as a "human ensemble." Such ensembles, we saw, can achieve remarkable predictive power if we let the right kind of aggregator learn to optimally combine the judgments of their members.

It is important to acknowledge the limitations of our study, the most notable one being due to the computational cost of our more advanced models. Our validation approach—using population bootstraps for simpler models and multi-seed runs for the neural networks—was a pragmatic response to the fact that a full bootstrap re-optimization would have kept our computer busy for weeks (or would have required spending a small fortune on Amazon's AWS). While our validation approach provides reliable uncertainty estimates for each model class, allowing for principled statistical comparisons between successive stages, we admit that a full bootstrap re-optimization would be required to validate the stability of the entire parameter-finding pipeline itself. This must be left for future work, hopefully to be carried out with access to greater computational resources.

Moreover, our findings are based on a single, albeit large, dataset of probabilistic judgments where the primary goal was to maximize predictive performance as measured by accuracy and Brier score. We have shown that under these conditions, a learned, non-linear aggregator can decisively outperform all other methods. But what the best aggregation strategy is may be context-dependent, varying with the nature of the task, the structure of the crowd, or the specific goals of the aggregation (e.g., Shinitzky et al., 2025). Future work could compare the aggregators considered in this paper on other datasets, and could also consider their use for other purposes (e.g., if the ultimate goal is to reach optimal group decisions).

Also, while our findings connect with a substantial literature on expert forecast aggregation in decision analysis, this literature has produced *numerous* sophisticated methods, including contribution-based weighting that estimates forecasters' marginal contributions to accuracy (Budescu and Chen, 2015), IRT-based models that jointly estimate item difficulty and forecaster ability (Bo et al., 2017), and regularized approaches to learning forecaster-specific parameters (Satopää, 2022). Our study is not intended as an exhaustive benchmark of all available aggregation methods. Rather, our goal was twofold: first, to introduce a new application of bounded-confidence opinion dynamics to probability aggregation, showing that the Hegselmann–Krause model—originally developed to model social influence—can serve as an effective aggregation mechanism; and second, to show that neural networks can substantially outperform both classical aggregation methods and our HK-based approach. Nevertheless, a systematic comparison across the full range of proposed aggregators (and then ideally, as just mentioned, also using multiple datasets) would be a fruitful project.

As a further avenue for future research, we mention that prediction markets offer a unique opportunity for real-world validation of our aggregation methods. Suppose that we still had access to the participants in the study reported in Stinson et al. (2025). Then we could try to recruit them again, now for forecasting tasks, and aggregate their predicted probabilities using our trained neural networks. In a next step, these aggregated forecasts could be tested against market prices on platforms such as Polymarket (https://polymarket.com/) or Metaculus (https://www.metaculus.com/). If trades based on these aggregated predictions were consistently or mostly profitable, that would be strong evidence that sophisticated aggregation can uncover information that even efficient markets miss.[11]

Our approach of improving collective performance by using increasingly more sophisticated aggregators is to be distinguished from, yet complementary to, another tested strategy for improving collective performance, viz., improving the quality of judgments through structured social interaction. For instance, Navajas et al. (2018) showed that allowing small subgroups to deliberate and form a consensus before averaging these consensus estimates also produces an aggregate that is superior to the traditional wisdom of large crowds. Note that their "procedural" solution improves the inputs to aggregation, while our "algorithmic" solution improves the aggregation function itself. An interesting avenue for future research would be to combine these approaches: first, to use deliberation to generate higher-quality initial judgments and then apply a learned aggregator, such as a neural network, to optimally combine those refined judgments.

In summary, our findings argue for a shift in perspective. Rather than asking *whether* crowds are wise, or *under what conditions* they are wise, we asked how their wisdom can be unlocked. Our results suggest that the limits of collective intelligence often lie not in the crowd itself but in the tools we use to listen to it. At least as far as our data go, we can conclude that, by recruiting more sophisticated aggregation methods, it is possible to progress through distinct stages of crowd wisdom, ultimately achieving a collective intelligence that goes beyond the abilities of any individual member.

Acknowledgments

We are indebted to Christopher von Bülow and two anonymous referees for valuable comments on previous versions.

ORCID iD

Igor Douven https://orcid.org/0000-0003-3413-080X

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Significance statement

When groups of people make predictions—about election outcomes, economic trends, or scientific questions—their combined

judgment is often surprisingly accurate, a phenomenon known as the "wisdom of the crowd." But how should we combine individual opinions into a collective prediction? Most research simply averages them. We show that this choice matters enormously. Using a dataset of 376 individuals' probability judgments on 1200 factual claims, we compared aggregation methods of increasing sophistication: simple averages, optimized weighted averages, a dynamic model of social learning, and neural networks. Performance improved at every step. Simple averages outperformed the typical individual but fell short of the crowd's best member. Optimized weights and a dynamic model based on bounded-confidence opinion dynamics—where individuals iteratively update their beliefs by attending to like-minded peers—pushed performance further. Neural networks, which learn an aggregation function directly from data, surpassed even the single best individual in the crowd, achieving a level of collective accuracy that no predefined formula could match. These findings reframe crowd wisdom not as a fixed property of a group but as something that can be unlocked to different degrees depending on the aggregation method employed. For basic science, this establishes a conceptual bridge between collective intelligence research and ensemble learning in machine learning, showing that human crowds function as "human ensembles" whose potential is constrained by how we listen to them.

Supplemental material

Supplemental material for this article is available online.

Supplementary Materials, containing the Julia code (Bezanson et al., 2017) used for the simulations and analyses we report, can be downloaded from this repository: https://osf.io/s9enj/overview?view_only=5db0093862a64e04ac4a018e36cb533f.

Notes

1. The HK model and various of its extensions have since been used to study a variety of phenomena, including peer disagreement (De Langhe, 2013; Douven, 2010), mis- and disinformation and opinion manipulation (Douven and Hegselmann, 2021; Glass and Glass, 2021; Hegselmann et al., 2015; Hegselmann and Krause, 2015), the social propagation of vague beliefs (Crosscombe and Lawry, 2016), various forms of non-deductive social reasoning (Douven, 2019, 2023; Douven and Schurz, 2024; Douven and Wenmackers, 2017; Trpin and Pellert, 2019), and bounded rationality (Lawry, 2024).
2. As Hegselmann and Krause (2005) do for the standard HK model.
3. For full details on data collection, experimental conditions, and participant demographics, see Stinson et al. (2025). Our present study and the Stinson et al. study address complementary questions using different analytical approaches. While Stinson and coauthors develop collective-inference algorithms based on judgment-generative models and individual calibration, evaluating performance under simulated sparsity conditions (resampling the data to vary the number of ratings per claim), we exploit the full dense matrix—where every participant rated every claim—to optimize opinion-dynamics models and train neural networks, asking whether these approaches can extract structure that simpler aggregation methods miss. There is no substantive overlap in the aggregation methods analyzed.
4. We used Borg MOEA because our accuracy objective is non-differentiable, involving a discontinuous threshold at 0.5. Julia's `BlackBoxOptim.jl` package provided robust support for this algorithm.
5. Indeed, participants' Brier scores and their accuracies are highly anti-correlated: $r = -.79$.
6. To forestall misunderstanding, we note that the two stability assessments quantify different sources of variability. The population bootstrap (used for the averaging and MHK models) captures sampling variability in the crowd by resampling participants, whereas the multi-seed protocol for neural networks captures optimization stochasticity (random initialization and minibatch order) on a fixed training set. Consequently, the uncertainty estimates to be reported are appropriate for their model class but should not be interpreted as identical estimators of the same underlying quantity.
7. The average and best individual here are the average and best individuals in the test set, though it was checked that the same result holds vis-à-vis the average and best individuals in the entire dataset. Also note that the best individual need not be—and, as we verified, in our case was not—the same participant for both criteria.
8. This finding is consistent with work on expert-weighting in forecasting tournaments; see Mellers et al. (2014) and Satopää et al. (2014).
9. A reasonable suggestion (which we owe to an anonymous referee) is that neural network performance could stem at least partly from implicit extremization, that is, learning to push aggregate probabilities away from .5 when forecasters share partially independent information. We tested this by implementing extremizing aggregators that apply a log-odds transformation with an optimized parameter to each base average (in the manner of Lichtendahl et al., 2017). While this did consistently improve Brier scores, these remained significantly higher than those obtained by the neural aggregators. Also, extremizing had no effect on accuracy. (This asymmetry in fact reflects the nature of the transformation: pushing a probability from .7 to .85 improves calibration if the event occurs, but both values yield the same binary classification.) This indicates that the networks learn patterns beyond simple extremization. For details of the analysis, see the Supplementary Materials.
10. The synergy between these fields is in fact bidirectional; just as we use deep learning to extract collective wisdom, others have begun incorporating principles from collective intelligence, such as self-organization and emergence, to design more robust and adaptable deep learning systems (Ha and Tang, 2022).
11. In actuality, we no longer have access to the participants from the Stinson et al. study. Hence, the validation as outlined here would require training the neural networks on a fresh batch of

participants, who would first have to complete a survey comparable to the one used by Stinson and colleagues and who would then have to be willing to return from time to time to post their predictions; such a validation would come with considerable expenses.

References

Amblard F and Deffuant G (2004) The role of network topology on extremism propagation with a bounded confidence model. *Physica A: Statistical Mechanics and Its Applications* 343: 725–738. https://doi.org/10.1016/j.physa.2004.06.102

Becker J, Porter M and Centola D (2019) The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences of the United States* 116: 10717–10722.

Bezanson J, Edelman A, Karpinski S, et al. (2017) Julia: a fresh approach to numerical computing. *SIAM Review* 59: 65–98. https://doi.org/10.1137/141000671

Bo YE, Budescu DV, Lewis C, et al. (2017) An IRT forecasting model: linking proper scoring rules to item response theory. *Judgment and Decision Making* 12: 90–103. https://doi.org/10.1017/S1930297500005647

Breiman L (1996) Bagging predictors. *Machine Learning* 24: 123–140. https://doi.org/10.1007/bf00058655

Budescu DV and Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Science* 61: 267–280. https://doi.org/10.1287/mnsc.2014.1909

Bullen PS (2003) *Handbook of Means and their Inequalities*. 2nd edition. Kluwer.

Chen S, Glass DH and McCartney M (2019) Two-dimensional opinion dynamics in social networks with conflicting beliefs. *AI & Society* 34: 695–704. https://doi.org/10.1007/s00146-017-0784-6

Crosscombe M and Lawry J (2016) A model of multi-agent consensus for vague and uncertain beliefs. *Adaptive Behavior* 24: 249–260. https://doi.org/10.1177/1059712316661395

Davis-Stober CP, Budescu DV, Dana J, et al. (2014) When is a crowd wise? *Decision* 1: 79–101. https://doi.org/10.1037/dec0000004

De Langhe R (2013) Peer disagreement under multiple epistemic constraints. *Synthese* 190: 2547–2556. https://doi.org/10.1007/s11229-012-0149-0

Deffuant G, Neau D, Amblard F, et al. (2000) Mixing beliefs among interacting agents. *Advances in Complex Systems* 3: 87–98. https://doi.org/10.1142/s0219525900000078

Douven I (2010) Simulating peer disagreements. *Studies In History and Philosophy of Science Part A* 41: 148–157. https://doi.org/10.1016/j.shpsa.2010.03.010

Douven I (2019) Optimizing group learning: an evolutionary computing approach. *Artificial Intelligence* 275: 235–251. https://doi.org/10.1016/j.artint.2019.06.002

Douven I (2023) Explaining the success of induction. *The British Journal for the Philosophy of Science* 72: 381–404. https://doi.org/10.1086/714796

Douven I and Hegselmann R (2021) Mis- and disinformation in a bounded confidence model. *Artificial Intelligence* 291: 103415. https://doi.org/10.1016/j.artint.2020.103415

Douven I and Hegselmann R (2022) Network effects in a bounded confidence model. *Studies In History and Philosophy of Science Part A* 94: 56–71. https://doi.org/10.1016/j.shpsa.2022.05.002

Douven I and Schurz G (2024) Integrating individual and social learning: Accuracy and evolutionary viability. *Computational & Mathematical Organization Theory* 30: 32–74. https://doi.org/10.1007/s10588-022-09372-1

Douven I and Wenmackers S (2017) Inference to the best explanation versus bayes' rule in a social setting. *The British Journal for the Philosophy of Science* 68: 535–570. https://doi.org/10.1093/bjps/axv025

Fortunato S, Latora V, Pluchino A, et al. (2005) Vector opinion dynamics in a bounded confidence consensus model. *International Journal of Modern Physics C* 16: 1535–1551. https://doi.org/10.1142/s0129183105008126

Fu G, Zhang W and Li Z (2015) Opinion dynamics of modified Hegselmann–Krause model in a group-based population with heterogeneous bounded confidence. *Physica A: Statistical Mechanics and Its Applications* 419: 558–565. https://doi.org/10.1016/j.physa.2014.10.045

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

Genest C and Zidek JV (1986) Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* 1: 114–135. https://doi.org/10.1214/ss/1177013825

Glass CA and Glass DH (2021) Opinion dynamics of social learning with a conflicting source. *Physica A: Statistical Mechanics and Its Applications* 563: 125480. https://doi.org/10.1016/j.physa.2020.125480

Ha D and Tang Y (2022) Collective intelligence for deep learning: a survey of recent developments. *Collective Intelligence* 1: 26339137221114874. https://doi.org/10.1177/26339137221114874

Hadka D and Reed P (2013) Borg: an auto-adaptive many-objective evolutionary computing framework. *Evolutionary Computation* 21: 231–259. https://doi.org/10.1162/EVCO_a_00075

Hegselmann R and Krause U (2002) Opinion dynamics and bounded confidence: models, analysis, and simulations. *The Journal of Artificial Societies and Social Simulation*. https://jasss.soc.surrey.ac.uk/5/3/2.html

Hegselmann R and Krause U (2005) Opinion dynamics driven by various ways of averaging. *Computational Economics* 25: 381–405. https://doi.org/10.1007/s10614-005-6296-3

Hegselmann R and Krause U (2015) Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: a simple unifying model. *Networks and Heterogeneous Media* 10: 477–509. https://doi.org/10.3934/nhm.2015.10.477

Hegselmann R, König S, Kurz S, et al. (2015) Optimal opinion control: the campaign problem. *The Journal of Artificial Societies and Social Simulation* 18: 18. https://doi.org/10.18564/jasss.2847

Hong L and Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States* 101: 16385–16389. https://doi.org/10.1073/pnas.0403723101

Hong L and Page SE (2025) The range of collective accuracy for binary classifications under majority rule. *Economic Theory* 79: 275–300. https://doi.org/10.1007/s00199-024-01570-z

Huang S, Golman R and Broomell SB (2024) Combining the aggregated forecasts: stacking multiple weighting models. *Decision Analysis* 21: 70–86.

Huet S, Deffuant G and Jager W (2008) A rejection mechanism in 2D bounded confidence provides more conformity. *Advances in Complex Systems* 11: 529–554. https://doi.org/10.1142/s0219525908001799

Landemore H (2012) *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press.

Lawry J (2024) Multiple belief states in social learning: an evidence tokens model. *Synthese* 204: 121. https://doi.org/10.1007/s11229-024-04770-1

Lichtendahl KC, Grushka-Cockayne Y, Jose VR, et al. (2017). Extremizing and anti-extremizing in Bayesian ensembles of binary-event forecasts. https://ssrn.com/abstract=2940740

Lorenz J (2007) Continuous opinion dynamics under bounded confidence: a survey. *International Journal of Modern Physics C* 18: 1819–1838. https://doi.org/10.1142/s0129183107011789

Lorenz J (2008) Fostering consensus in multidimensional continuous opinion dynamics under bounded confidence. In: Helbing D (ed) *Managing Complexity: Insights, Concepts, Applications*. Springer, pp. 321–334. https://doi.org/10.1007/978-3-540-75261-5_15

Lorenz J, Rauhut H, Schweitzer F, et al. (2011) How social influence can undermine the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States* 108: 9020–9025. https://doi.org/10.1073/pnas.1008636108

Mellers B, Ungar L, Baron J, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science* 25: 1106–1115. https://doi.org/10.1177/0956797614524255

Morris PA (1983) An axiomatic approach to expert resolution. *Management Science* 29: 24–32. https://doi.org/10.1287/mnsc.29.1.24

Navajas J, Niella T, Garbulsky G, et al. (2018) Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* 2: 126–132. https://doi.org/10.1038/s41562-017-0273-4

Page SE (2007) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Prelec D, Seung HS and McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541: 532–535. https://doi.org/10.1038/nature21054

Satopää VA (2022) Regularized aggregation of one-off probability predictions. *Operations Research* 70: 3558–3580. https://doi.org/10.1287/opre.2021.2224

Satopää VA, Baron J, Foster DP, et al. (2014) Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting* 30: 344–356. https://doi.org/10.1016/j.ijforecast.2013.09.009

Schoenegger P, Tuminauskaite N, Park S, et al. (2024) Wisdom of the silicon crowd: large language model ensembles match human collective forecasting accuracy. *Science Advances* 10: eadk2543.

Shinitzky H, Parpara O, Ezrets M, et al. (2025) The meta-aggregation approaches: correctly choosing how to choose correctly in groups. *Collective Intelligence* 4: 26339137251315410. https://doi.org/10.1177/26339137251315410

Stinson P, van den Bosch J, Jerde T, et al. (2025) Collective inference of the truth of propositions from crowd probability judgments. arXiv 2501.04983. https://arxiv.org/abs/2501.04983

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Random House.

Trpin B and Pellert M (2019) Inference to the best explanation in uncertain evidential situations. *The British Journal for the Philosophy of Science* 70: 977–1001. https://doi.org/10.1093/bjps/axy027