The Power of Meta-Prediction Accuracy: Leveraging Predictions of Others' Predictions to Enhance Collective and Individual Intelligence

YUNHAO ZHANG, University of California, Berkeley, USA
EAMAN JAHANI, University of Maryland, USA
DOUGLAS GUILBEAULT, Stanford University, USA
JULIANA SCHROEDER, University of California, Berkeley, USA

Human cultural evolution and team performance is driven by collective intelligence — the ability for individuals to improve the accuracy of their beliefs by learning from others with diverse opinions. But a continual challenge that individuals face is how to identify and learn from reliable peers in the absence of objective performance measures indicating each other's accuracy. Here, we propose that one cue people can use to ascertain others' accuracy – even when the ground truth is unavailable – is others' meta-prediction accuracy (i.e., how well they can predict others' predictions). Building on a theoretical link between prediction accuracy and meta-prediction accuracy, we prove that, for probabilistic prediction problems (e.g., estimating the likelihood that a statement is true), aggregating over the predictions from the group of agents with better average meta-prediction accuracy yields asymptotically accurate predictions. Empirical validation across various epistemic contexts confirms that our simple aggregation rule outperforms traditional wisdom-of-the-crowds aggregation methods (e.g. majority vote, average probability, and confidence-based aggregation), and performs on par with, or better than, state-of-the-art techniques (e.g., Prelec et al. 2017; Palley & Soll 2019; Palley & Satopää 2023). Crucially, our method is both psychologically intuitive and easily explainable. Our social influence experiment demonstrates that people are preferentially influenced by aggregate predictions from the group agents with better average meta-prediction accuracy. Because meta-prediction accuracy is not only socially influential but also reliably diagnostic of accuracy, it offers a powerful yet novel mechanism for improving social learning – one that outperforms traditional cues such as confidence or social proof (i.e., what the majority believes).

CCS Concepts: • **Wisdom of Crowds**; • **Social Influence**; • **Meta-prediction**;

Additional Key Words and Phrases: Social Learning, Explainable Algorithm

ACM Reference Format:
Yunhao Zhang, Eaman Jahani, Douglas Guilbeault, and Juliana Schroeder. 2018. The Power of Meta-Prediction Accuracy: Leveraging Predictions of Others' Predictions to Enhance Collective and Individual Intelligence. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym CI'25).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

One of the fundamental questions in the field of collective intelligence is how to improve individual judgment through social influence. Although research has shown that using conventional wisdom-of-crowds estimates (e.g., majority

---

Authors' Contact Information: Yunhao Zhang, yunhao.jerry.zhang@gmail.com, University of California, Berkeley, Berkeley, California, USA; Eaman Jahani, eaman@umd.edu, University of Maryland, College Park, Maryland, USA; Douglas Guilbeault, Stanford University, Palo Alto, California, USA, dguilb@stanford.edu; Juliana Schroeder, University of California, Berkeley, Berkeley, California, USA.

---

vote) as social influence may improve individual judgment (1, 3, 4, 6, 19), the average improvement is modest, because the signals are sometimes inaccurate (9, 10, 14, 15, 17). On the other hand, recent advancement in Wisdom of Crowds methods can reliably produce more accurate signals by eliciting agents' meta-predictions by asking agents to provide not only their predictions of a real-world outcome, but also their meta-predictions – their predictions of other agents' predictions (e.g., the estimated proportion of advisors providing a particular response or the average numerical prediction made by other agents) (18, 22–25, 28). Despite their potential, the existing meta-prediction methods rely on complex mathematical modeling to justify their validity, and sophisticated algorithms to execute, which makes them difficult to communicate and reason about. This limits their usage among individual decision-makers who tend to favor intuitive and explainable procedures (2, 5, 7, 29, 30), which is crucial to facilitate social learning (4).

The current paper addresses this gap by developing a wisdom-of-crowds method that uses meta-prediction accuracy to produce a more accurate aggregation compared to conventional methods such as group average, majority opinion, and confidence-based aggregation, while remaining easy to explain and communicate as a social influence tool when used in a social learning context.

2 Our Aggregation Method

In binary classification problems, the following procedure can better harness the Wisdom of Crowds:

(1) Elicit agents' probabilistic predictions of a binary event (e.g., what is the probability of the statement being True?). Suppose $Z_i \in [0,1]$ corresponds to agent i's individual prediction. We can convert this probabilistic prediction to a binary vote, X_i, using a threshold of 0.5. A probability prediction less than 50% is classified as a "False" vote and a probability greater than 50% is classified as a "True" vote:

$$X_i = \begin{cases} 1 & \text{if } Z_i > 0.5 \\ 0 & \text{if } Z_i < 0.5 \\ 0 \text{ or } 1 & \text{if } Z_i = 0.5 \end{cases} \tag{1}$$

Optionally, one may directly elicit the binary vote X_i to avoid randomly classifying agents with probability exactly equal to 50%.

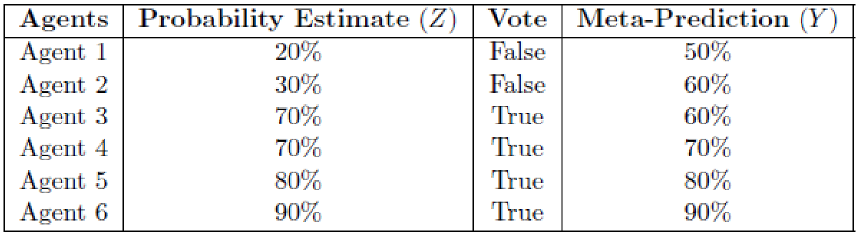
(2) Elicit agents' meta-predictions (e.g., what is the average probability estimated by other forecasters?)[1]. Let $Y_i \in [0,1]$ correspond to the agent i's meta-prediction.
(3) Compute the average probabilistic prediction (from Step 1) of all agents, which is the ground truth for meta-prediction.
(4) Compute the average meta-prediction among those with a "False" vote and the average meta-prediction among those with a "True" vote, respectively, and calculate the absolute error of each group's average meta-prediction with respect to the ground truth for meta-prediction.
(5) The response category containing agents with the more accurate (average) meta-prediction is the *binary* output of our aggregation rule.

---

[1] One may also ask "what is the average probability estimated by all forecasters?", because the ways of asking elicit the same underlying information, since all agents know their own probability.

(6) (Additionally) The average probabilistic prediction of agents in the response category described in Step 5 is the *probabilistic* output of our aggregation rule.

As an illustration, suppose there are six agents with their responses to the prediction and meta-prediction questions listed in Table 1. Agent 1 and Agent 2 believe the statement is false (i.e., less than 50% probability of the statement being true), and the average meta-prediction of these two agents is 55%. The remaining four agents believe the statement is "true", and the average meta-prediction of these four agents is 75%. Since the average probability estimate of the entire group of six agents is 60%, the two agents who think the statement is false more accurately predict the group mean compared to those who think the statement is true. Therefore, "false" is the binary output based on the meta-prediction-based aggregation rule, and "25%" is its probabilistic output.



Agents	Probability Estimate (Z)	Vote	Meta-Prediction (Y)
Agent 1	20%	False	50%
Agent 2	30%	False	60%
Agent 3	70%	True	60%
Agent 4	70%	True	70%
Agent 5	80%	True	80%
Agent 6	90%	True	90%

Table 1: *Agents' Responses: (1) probability of a statement being true; (2) average probability estimated by other forecasters*

We can prove the following proposition for our aggregation rule (see Appendix): for any binary classification problem in which there exists a positive correlation between binary prediction accuracy and meta-prediction accuracy, our proposed aggregation rule would select the correct response, regardless of whether the majority is correct or not. Therefore in large groups, social influence signals leveraging meta-prediction accuracy is a strict improvement over signals based on majority prediction.

3 Empirical Analysis: Tests of the Association between Prediction Accuracy and Meta-prediction Accuracy, and the Aggregation Rule

3.1 Context 1: Questions with Intuitive versus Counter-intuitive Truth

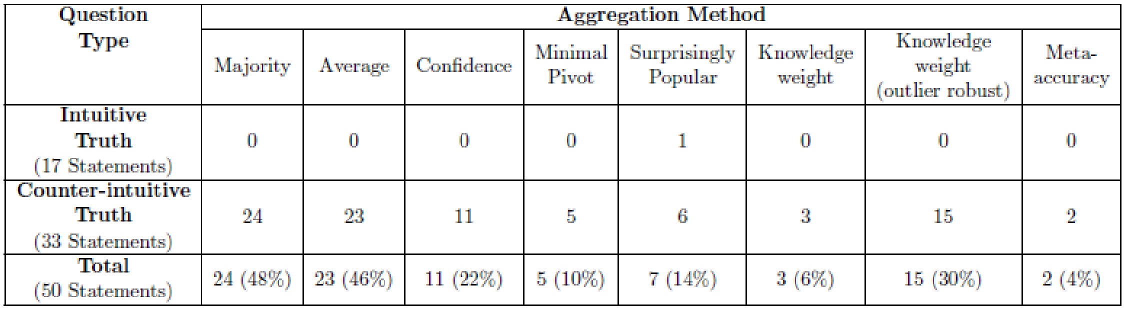
In their Study 1, Wilkening et al. (2022) collected 89 participants' responses to fifty True or False statements regarding whether the largest city in a state is the state capital for each of the fifty states in the United States. A notable feature of this context is that the largest city is not necessarily the state capital, contrary to many people's intuition. For example, Chicago, the largest city in Illinois, is not its state capital. The largest city is the state capital in only 17 of 50 states.

For example, one of the question statements is "Chicago is the state capital of Illinois." For each statement, the researchers asked the participants:

- *Question 1. Is this statement more likely to be true or false?*
- *Question 2. What percentage of other people do you think thought the statement was true?*
- *Question 3. What is the probability that the statement is true?*
- *Question 4. What is the average probability estimated by the other forecasters?*

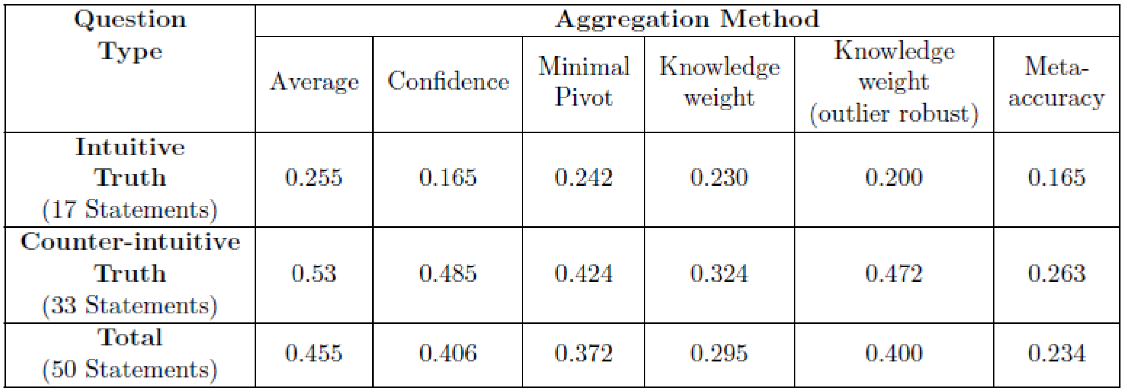
3.1.1 Context 1 Results: Association between Prediction Accuracy and Meta-Prediction Accuracy. We regress prediction accuracy from Question 1 (incorrect answers are coded as 1 and correct answers as 0) on meta-prediction accuracy from Question 4 (i.e., the absolute error of one's estimate of the average probability estimate of other respondents in the sample) with question and participant fixed effects, and robust standard errors clustered at the participant level. We find a strong positive association between prediction accuracy and meta-prediction accuracy ($b = 0.814$, $t = 7.2$, 95% CI of b = [0.59, 1.04], $p < 0.0001$). In addition, this positive association holds across intuitive and counter-intuitive questions.

3.1.2 Context 1 Results: Meta-Accuracy-based Aggregation Method. As shown in Table 2, our method is the most accurate (2 inaccuracies out of 50 statements), compared to the majority rule, the average probabilistic estimate (i.e., 50% as a cutoff for classification), the group-confidence-based method, the Surprisingly Popular Algorithm (25), the minimal pivot method (22), and the knowledge-weighted method (23) which are incorrect in 24, 23, 11, 5, 7, 3 and 15 of the 50 questions, respectively. Our algorithm achieves the lowest RMSE when comparing the performance based on Root Mean Squared Error (RMSE)(see Table 3).



Question Type	Aggregation Method							
	Majority	Average	Confidence	Minimal Pivot	Surprisingly Popular	Knowledge weight	Knowledge weight (outlier robust)	Meta-accuracy
Intuitive Truth (17 Statements)	0	0	0	0	1	0	0	0
Counter-intuitive Truth (33 Statements)	24	23	11	5	6	3	15	2
Total (50 Statements)	24 (48%)	23 (46%)	11 (22%)	5 (10%)	7 (14%)	3 (6%)	15 (30%)	2 (4%)

Table 2: *Number of inaccurate answers computed using each aggregation method among statements where the truth is intuitive (17 statements) versus counter-intuitive (33 statements).*



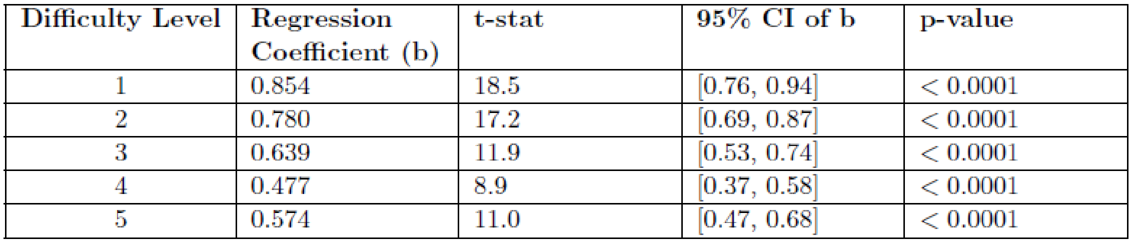
Question Type	Aggregation Method					
	Average	Confidence	Minimal Pivot	Knowledge weight	Knowledge weight (outlier robust)	Meta-accuracy
Intuitive Truth (17 Statements)	0.255	0.165	0.242	0.230	0.200	0.165
Counter-intuitive Truth (33 Statements)	0.53	0.485	0.424	0.324	0.472	0.263
Total (50 Statements)	0.455	0.406	0.372	0.295	0.400	0.234

Table 3: *Root Mean Squared Error (RMSE) computed using each aggregation method among statements where the truth is intuitive (17 statements) versus counter-intuitive (33 statements).*

3.2 Context 2: Questions with varying Difficulty Levels

Wilkening et al. (2022) also generated 500 true or false scientific statements, separated into five levels of difficulty with 100 problems for each difficulty level.

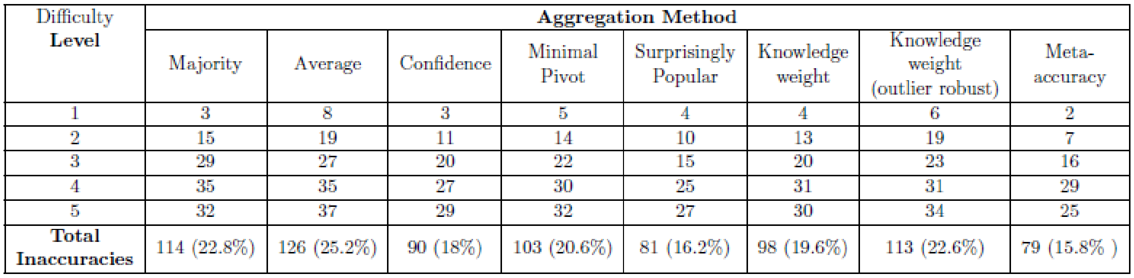
3.2.1 Context 2 Results: Association between Prediction Accuracy and Meta-Prediction Accuracy. Using the same regression as in Context 1, we find a positive association between prediction accuracy and meta-prediction accuracy for all difficulty levels (see Table 4).



Difficulty Level	**Regression Coefficient (b)**	**t-stat**	**95% CI of b**	**p-value**
1	0.854	18.5	[0.76, 0.94]	< 0.0001
2	0.780	17.2	[0.69, 0.87]	< 0.0001
3	0.639	11.9	[0.53, 0.74]	< 0.0001
4	0.477	8.9	[0.37, 0.58]	< 0.0001
5	0.574	11.0	[0.47, 0.68]	< 0.0001

Table 4: *Regression coefficient (b) characterizes the association between meta-prediction accuracy and prediction accuracy at each difficulty level.*

3.2.2 Context 2 Results: Meta-Accuracy-based Aggregation Method . As summarized in Table 5, our meta-accuracy-based method outperforms conventional methods (e.g. group mean, majority vote, confidence-based aggregation) across difficulty levels, and is on par than the advanced methods. This conclusion holds for aggregate outcomes measured by RMSE (Table 6).



Difficulty **Level**	**Aggregation Method**							
	Majority	Average	Confidence	Minimal Pivot	Surprisingly Popular	Knowledge weight	Knowledge weight (outlier robust)	Meta-accuracy
1	3	8	3	5	4	4	6	2
2	15	19	11	14	10	13	19	7
3	29	27	20	22	15	20	23	16
4	35	35	27	30	25	31	31	29
5	32	37	29	32	27	30	34	25
Total Inaccuracies	114 (22.8%)	126 (25.2%)	90 (18%)	103 (20.6%)	81 (16.2%)	98 (19.6%)	113 (22.6%)	79 (15.8%)

Table 5: *Number of inaccurate answers computed using each aggregation method at each difficulty level. There are 100 questions in total for each difficulty level. The performance comparisons pooling all 500 questions are reported in the main text.*

4 The Social Influence Experiment

This experiment (N = ~ 800) demonstrates the applicability of our algorithm in the context of social learning. Full experimental materials are available on our OSF site: https://osf.io/85qfv/?view_only=9b78be49015f4b6687add07aa8bd02d7

Difficulty Level	Aggregation Method					
	Average	Confidence	Minimal Pivot	Knowledge weight	Knowledge weight (outlier robust)	Meta-accuracy
1	0.319	0.238	0.274	0.260	0.268	0.226
2	0.385	0.339	0.343	0.326	0.347	0.293
3	0.428	0.394	0.394	0.377	0.395	0.371
4	0.459	0.446	0.439	0.430	0.441	0.454
5	0.460	0.462	0.440	0.431	0.447	0.436
RMSE	0.414	0.385	0.383	0.371	0.386	0.366

Table 6: *Average Root Mean Squared Error (RMSE) computed using each aggregation method at each difficulty level. There are 100 questions in total for each difficulty level. The performance comparisons pooling all 500 questions are reported in the main text.*

4.0.1 Procedure. We adopted the fifty statements used in Wilkening et al. (2022) about whether a city in a state is the state capital. For each of the fifty statements, we first elicited participants' prior probabilistic judgment about the statement (i.e. "*what is the probability that this statement is 'True'?*") Then, in a between-subjects design, we randomly assigned participants into one of four social influence conditions: the majority influence condition, the confidence influence condition, the meta-accuracy influence condition, and the meta-accuracy framed as a random influence condition. All of the social influence information was generated based on participants' responses in Wilkening et al. (2022). Each participant received the same type of social influence message across all fifty statements. After presenting one of the social influence messages, we elicited participants' posterior probabilistic judgment about the statement ("*Now what do you think is the probability that the statement is 'True'?*"). The exact social influence messages provided to the participants are included below:

1. **Majority Influence**: In this condition, participants read: "*In the previous group of participants, the majority thought the statement was '[True / False]'*".
2. **Confidence Influence**: In this condition, participants read: "*......, the people who said '[True / False]' reported more confidence in their answer than those who said '[False / True]'*".
3. **Meta-Accuracy Influence**: In this condition, participants read: "*......, the people who said '[True / False]' could more accurately predict the average of other people's estimates than those who said '[False / True]'*".

Finally, one may be concerned that just because participants update toward the meta-accuracy signal does not mean they view it as an interpretable and informative signal of prediction accuracy, since they could be blindly following whatever signal they are presented with. To explore this possible confound, we include a "coin influence" condition. In this condition, we provided participants with information about the result of a "random coin clip": "*We flipped a coin with an equal chance of landing on heads or tails, assuming heads means the statement is 'True' and tails means it is 'False'. The coin came up [heads / tails], suggesting the statement is '[True / False]*". Crucially, unbeknownst to the participants, the outcome of the coin flip was matched to the Meta-Accuracy Influence. Then if participants indeed think the communication in the meta-accuracy influence is uninformative or uninterpretable, we should expect that the extent of updating in the meta-accuracy and the coin-flip influence conditions are the same.

4.1 Results

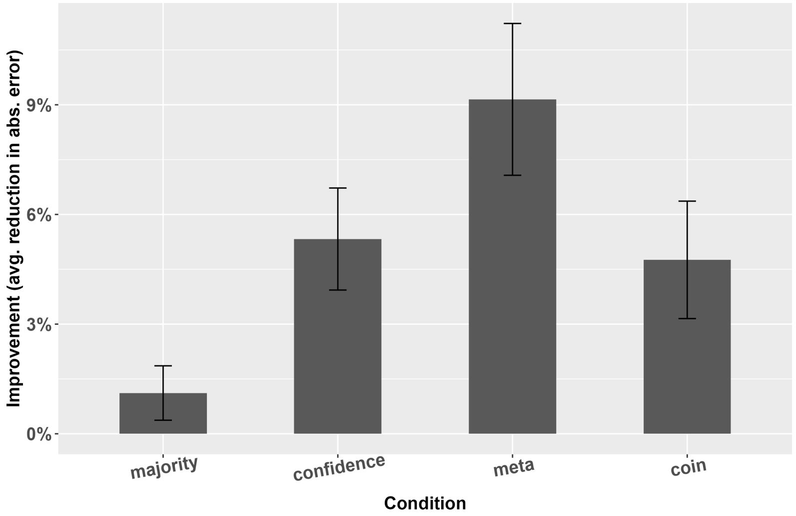


Fig. 1. *The x-axis is the four types of social influence conditions – signals generated by advisors' majority opinion, advisors' confidence levels, advisors' meta-accuracy levels, and advisors' meta-accuracy levels but framed as a random coin flip (aka "coin influence"). The y-axis is the average improvement in participants' probability estimates from before versus after they received the social influence information. The bars indicate 95% confidence intervals based on robust standard errors.*

As one may observe from Figure, meta-accuracy-based influence on average resulted in a larger improvement compared to the other types of social influence. In addition, although we are not arguing that people completely follow the logic proposed in the theoretical model, the comparison between meta-accuracy influence and the "coin" influence suggests people indeed treat better meta-prediction accuracy as an informative signal of prediction accuracy (compared to a random process benchmark).

5 Conclusion

Our paper makes two advances. First, leveraging the empirical phenomenon that people with more accurate predictions tend to have more accurate meta-predictions, we develop a novel aggregation procedure that can be used in binary classification problems. We then empirically confirm that our proposed aggregation rule outperforms conventional crowd wisdom estimates such as majority vote, group mean, and confidence-based aggregation, while remaining equally accurate as (and in some cases, more accurate than) the advanced Surprisingly Popular algorithm (25) the Minimal Pivoting method (22), the knowledge-weighted method (23). Second, we show that individual decision-makers can and do incorporate advisors' meta-prediction accuracy into their own predictions, substantially improving their predictions through belief-updating in a social influence context. The results hold because (1) signals based on meta-prediction accuracy are more likely to be accurate, which allows individual decision makers to update towards the truth, and (2) individual decision makers naturally *perceive* advisors with better meta-prediction accuracy as having better prediction accuracy, and thus are willing to rely on advisors' meta-prediction accuracy to update.

References

[1] Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., and Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21), 11379-11386.

[2] Bauer, K., von Zahn, M., and Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 34(4), 1582-1602.

[3] Becker, J., Brackbill, D., and Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26), E5070-E5076.

[4] Becker, J. A., Guilbeault, D., and Smith, E. B. (2022). The crowd classification problem: Social dynamics of binary-choice accuracy. *Management Science*, 68(5), 3949-3965.

[5] Bonezzi, A., Ostinelli, M., and Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General*, 151(9), 2250.

[6] Centola, D., Becker, J., Zhang, J., Aysola, J., Guilbeault, D., and Khoong, E. (2023). Experimental evidence for structured information–sharing networks reducing medical errors. *Proceedings of the National Academy of Sciences*, 120(31), e2108290120.

[7] Cadario, R., Longoni, C., and Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, 5(12), 1636-1642.

[8] Charness, G., and Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104, 1-23.

[9] Csaszar, F. A., and Eggers, J. P. (2013). Organizational decision making: An information aggregation view. *Management Science*, 59(10), 2257-2277.

[10] Davis-Stober, C. P., Budescu, D. V., Dana, J., and Broomell, S. B. (2014). When is a crowd wise?. *Decision*, 1(2), 79.

[11] DellaVigna, S., and Pope, D. (2018). Predicting experimental results: who knows what?. *Journal of Political Economy*, 126(6), 2410-2456.

[12] Fox, C. R., and Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Kirkebøen, and H. Montgomery (Eds.), *Essays in Judgment and Decision Making*. Oslo: Universitetsforlaget.

[13] Frey, V., and Van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, 67(7), 4273-4286.

[14] Galesic, M., Barkoczi, D., and Katsikopoulos, K. (2015). Wisdom of Randomly Assembled Small Crowds. *In Proceedings of the Annual Meeting of the Cognitive Science Society*, 37.

[15] Galesic, M., Barkoczi, D., and Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, 5(1).

[16] Guilbeault, D., and Centola, D. (2020). Networked collective intelligence improves dissemination of scientific information regarding smoking risks. *PLOS One*, 15(2), e0227813.

[17] Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, 107(2), 276.

[18] McCoy, J., and Prelec, D. (2023). A Bayesian Hierarchical Model of Crowd Wisdom Based on Predicting Opinions of Others. *Management Science*.

[19] Minson, J. A., Mueller, J. S., and Larrick, R. P. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science*, 64(9), 4177-4192.

[20] Moore, D. A., and Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, 11(8), e12331.

[21] Moore, D. A. (2022). Overprecision is a property of thinking systems. *Psychological Review*.

[22] Palley, A. B., and Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, 65(5), 2291-2309.

[23] Palley, A. B., and Satopää, V. A. (2023). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. *Management Science*, 69(9), 5128-5146.

[24] Peker, C. (2023). Extracting the collective wisdom in probabilistic judgments. *Theory and Decision*, 94(3), 467-501.

[25] Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532-535.

[26] Rader, C. A., Larrick, R. P., and Soll, J. B. (2017). Advice as a form of social influence: Informational motives and the consequences for accuracy. *Social and Personality Psychology Compass*, 11(8), e12329.

[27] Schmidt, F. L., and Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262.

[28] Wilkening, T., Martinie, M., and Howe, P. D. (2022). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. *Management Science*, 68(1), 487-508.

[29] Saranya, A., and Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7, 100230.

[30] Senoner, J., Netland, T., and Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. *Management Science*, 68(8), 5704-5723.

A Proof of Proposition 1

A.1 Statistical Property of the Aggregation Rule

In the procedure above, the true value of Y is known given the distribution of individual assessments Z_i. Thus the aggregate quantity $\mathbb{E}(\text{error in } Y|X = x)$ can be easily measured for each value of individual binary response classification $x \in \{0, 1\}$. This quantity indicates which individual assessment, x, is likely to be more accurate in meta-prediction. A desirable property is that the group with the correct (binary) prediction is more likely to have a more accurate meta-prediction:

$$\mathbb{E}(\text{error in } Y|X \text{ is correct}) < \mathbb{E}(\text{error in } Y|X \text{ is incorrect})$$

We can show this property holds as long as the accuracy in binary prediction is positively correlated with accuracy in meta-prediction (or accuracy in binary prediction is negatively correlated with error in meta-prediction). Before doing so, we define two new variables:

(1) **Classification Accuracy** is an indicator variable whether the individual binary prediction of agent i is correct:

$$X_i^a = \begin{cases} 1 & \text{if } X_i \text{ is correct,} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

(2) **Meta-Prediction Error** is a continuous variable measuring the error in the meta-prediction of agent i:

$$Y_i^a = Y_i - Y_{\text{true}} \tag{3}$$

where $Y_{\text{true}} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Proposition 1. *If* $cov(X^a, Y^a) < 0$*, then we have:*

$$\mathbb{E}(Y^a|X^a = 1) < \mathbb{E}(Y^a) < \mathbb{E}(Y^a|X^a = 0)$$

Proof. The proof follows from simple identities of covariance:

$$\begin{aligned}
&\text{cov}(X^a, Y^a) < 0 \\
\Rightarrow &\mathbb{E}[X^a Y^a] - \mathbb{E}[X^a]\mathbb{E}[Y^a] < 0 \\
\Rightarrow &\mathbb{E}(1 \times Y^a|X^a = 1)\mathbb{P}(X^a = 1) + \mathbb{E}(0 \times Y^a|X^a = 0)\mathbb{P}(X^a = 0) - \mathbb{P}(X^a = 1)\mathbb{E}(Y^a) < 0 \\
\Rightarrow &\mathbb{E}(Y^a|X^a = 1)\mathbb{P}(X^a = 1) - \mathbb{P}(X^a = 1)\mathbb{E}(Y^a) < 0 \\
\Rightarrow &\mathbb{E}(Y^a|X^a = 1) < \mathbb{E}(Y^a)
\end{aligned}$$

The other part of the inequality can be obtained by adding and subtracting $\mathbb{E}[Y^a]$ term.

$$
\begin{aligned}
&\mathrm{cov}(X^a, Y^a) < 0\\
\Rightarrow&\mathbb{E}[X^aY^a] - \mathbb{E}[X^a]\mathbb{E}[Y^a] < 0\\
\Rightarrow&\mathbb{E}[X^aY^a] - \mathbb{E}[Y^a] - \mathbb{E}[X^a]\mathbb{E}[Y^a] + \mathbb{E}[Y^a] < 0\\
\Rightarrow&\mathbb{E}[1 - X^a]\mathbb{E}[Y^a] - \mathbb{E}[(1 - X^a)Y^a] < 0\\
\Rightarrow&\mathbb{P}(X^a = 0)\mathbb{E}(Y^a) - \mathbb{E}(0 \times Y^a|X^a = 1)\mathbb{P}(X^a = 1) - \mathbb{E}(1 \times Y^a|X^a = 0)\mathbb{P}(X^a = 0) < 0\\
\Rightarrow&\mathbb{E}(Y^a|X^a = 0)\mathbb{P}(X^a = 0) - \mathbb{P}(X^a = 0)\mathbb{E}(Y^a) > 0\\
\Rightarrow&\mathbb{E}(Y^a|X^a = 0) > \mathbb{E}(Y^a)
\end{aligned}
$$

By combining the inequalities we obtain the desired result as stated in Proposition 1. □