Supplementary Information

August 18, 2025

1 A Proof of Improvement of Our Algorithm Based on Meta-prediction Accuracy Over Majority Vote

We prove that for any binary classification problem in which there exists a positive correlation between binary prediction accuracy and meta-prediction accuracy (or accuracy in binary prediction is negatively correlated with error in meta-prediction), our proposed aggregation algorithm would select the correct response, regardless of whether the majority is correct or not. Denote Y as the average probabilistic prediction of the group. The true value of Y is known given the distribution of individual probabilistic predictions Z_i. Thus the aggregate quantity $\mathbb{E}(\text{error in } Y \,\big|\, X = x)$ can be easily measured for each value of individual binary response classification (X) and $x \in \{0, 1\}$. 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:

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

Before proving this property, we define two new variables:

1. **Classification Accuracy** is an indicator variable whether the individual binary prediction of agent i is correct:
$$X_i^a = \begin{cases} 1 & \text{if } X_i \text{ is correct,} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

2. **Meta-Prediction Error** is a continuous variable measuring the error in the meta-prediction of agent i:
$$Y_i^a = Y_i - Y_{\text{true}} \tag{2}$$
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:*

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

Proof. The proof follows from simple identities of covariance:

$$\begin{aligned}
&\operatorname{cov}(X^a, Y^a) < 0 \\
\Rightarrow &\mathbb{E}[X^a Y^a] - \mathbb{E}[X^a]\mathbb{E}[Y^a] < 0 \\
\Rightarrow &\mathbb{E}(1 \times Y^a \,\big|\, X^a = 1)\mathbb{P}(X^a = 1) + \mathbb{E}(0 \times Y^a \,\big|\, X^a = 0)\mathbb{P}(X^a = 0) - \mathbb{P}(X^a = 1)\mathbb{E}(Y^a) < 0 \\
\Rightarrow &\mathbb{E}(Y^a \,\big|\, X^a = 1)\mathbb{P}(X^a = 1) - \mathbb{P}(X^a = 1)\mathbb{E}(Y^a) < 0 \\
\Rightarrow &\mathbb{E}(Y^a \,\big|\, X^a = 1) < \mathbb{E}(Y^a) \qquad (3)
\end{aligned}$$

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

$$\begin{aligned}
&\operatorname{cov}(X^a, Y^a) < 0 \\
\Rightarrow &\mathbb{E}[X^a Y^a] - \mathbb{E}[X^a]\mathbb{E}[Y^a] < 0 \\
\Rightarrow &\mathbb{E}[X^a Y^a] - \mathbb{E}[Y^a] - \mathbb{E}[X^a]\mathbb{E}[Y^a] + \mathbb{E}[Y^a] < 0 \\
\Rightarrow &\mathbb{E}[1 - X^a]\mathbb{E}[Y^a] - \mathbb{E}[(1 - X^a)Y^a] < 0 \\
\Rightarrow &\mathbb{P}(X^a = 0)\mathbb{E}(Y^a) - \mathbb{E}(0 \times Y^a \,\big|\, X^a = 1)\mathbb{P}(X^a = 1) - \mathbb{E}(1 \times Y^a \,\big|\, X^a = 0)\mathbb{P}(X^a = 0) < 0 \\
\Rightarrow &\mathbb{E}(Y^a \,\big|\, X^a = 0)\mathbb{P}(X^a = 0) - \mathbb{P}(X^a = 0)\mathbb{E}(Y^a) > 0 \\
\Rightarrow &\mathbb{E}(Y^a \,\big|\, X^a = 0) > \mathbb{E}(Y^a) \qquad (4)
\end{aligned}$$

By combining the inequalities 3 and 4 we obtain the desired result as stated in Proposition 1. □

2 The Positive Association Between Prediction Accuracy and Meta-prediction Accuracy

Using archival data from Wilkening et al. (2022), we test whether the hypothesized positive association between prediction accuracy and meta-prediction accuracy is robust, and whether our aggregation method outperforms conventional methods.

Their first dataset (a.k.a. the 50-state dataset; see SI Section 2.1) features 50 decision problems in which conventional methods tend to perform well (e.g., when the truth is intuitive and the majority tends to be correct) and poorly (e.g., when the truth is counterintuitive and the majority tends to be wrong while feeling confident). In particular, 89 participants provided 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.

Their second dataset (a.k.a. the 500-science-statement dataset; see SI Section 2.2) features 500 decision problems that vary in analytic difficulty (i.e., problems that are perceived to be difficult, as opposed to containing an undetected "lure" such that some people are confidently wrong). In particular, the authors generated 500 true or false statements regarding scientific facts that pertain to expected knowledge at a US primary and secondary grade school level.

For each statement, the researchers asked the participants the following questions:

- *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?*

In all of the analyses to follow, we examine prediction accuracy using the error of a response to a binary question. Incorrect responses, such as stating "False" for a true statement or "True" for a false statement, are coded as 1 and correct responses as 0. To measure meta-prediction accuracy, we use the absolute difference between an individual's meta-prediction and the average prediction of the other respondents in the group, whereby a lower difference indicates less error and thus greater accuracy. All materials, data, and analysis code are available on our OSF site: https://osf.io/85qfv/?view_only=983747da6e384227984f4291d2ff3be7

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

We examined the relationship between prediction accuracy and meta-prediction accuracy based on this data, and compared our aggregation method based on meta-prediction accuracy to existing benchmark methods. Using Question 3, we can calculate a measure of self-reported confidence in their binary response being correct, which is $50\% + |\textit{stated probability} - 50\%|$. This would allow us to examine the relationship between confidence and accuracy.

2.1.1 Context 1 Results: Association between Prediction Accuracy and Meta-Prediction Accuracy

To avoid multiple hypothesis testing (e.g., calculating Pearson's correlation for each statement), we regress prediction accuracy from Question 1[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. Since each participant answers multiple questions, robust standard errors clustered at the participant level are used for all regression results. All reported regression coefficients are raw and unstandardized throughout the paper, unless specified otherwise. 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], $p < 0.0001$).[2] This means that a one percentage point increase in the accuracy of meta-prediction is associated with 0.81 percentage point increase in the likelihood of correctly evaluating the veracity of the statement. We provide a summary of the Pearson's correlation statement-by-statement in SI Section 2.1.2, and show that all key findings are robust.

This positive association holds across intuitive and counter-intuitive questions. In the 17 states where the largest city is the state capital, the majority of participants correctly evaluated the statements. The regression coefficient for these statements is $b = 1.02$ ($t = 5.8$, 95% CI = [0.68, 1.36], $p < 0.0001$). For the 33 states where the largest city is not the state capital, and the truth is potentially counter-intuitive, the regression coefficient is $b = 0.572$ ($t = 4.1$, 95% CI = [0.30, 0.84], $p < 0.0001$).

In contrast, when using participants' self-reported confidence (calculated based on Question 3) as the independent variable, the regression coefficients for intuitive and counter-intuitive questions are $b = -0.76$ ($t = -5.4$, 95% CI of b = [-1.03, -0.48], $p < 0.0001$) and $b = 0.055$ ($t = 0.37$, 95% CI of b = [-0.24, 0.35], $p = 0.72$). Although confidence predicts accuracy when the correct answer is intuitive (higher confidence is associated with a lower likelihood of being incorrect), confidence fails to predict response accuracy when correct answers are counter-intuitive, suggesting that at least

[1] We use Question 1 instead of Question 2 to avoid randomly classifying those who stated 50% into "True" or "False". All conclusions are robust to using Question 2 for binary group classification.

[2] Adding participant fixed effects removes the cross-participant correlation between prediction accuracy and meta-prediction accuracy. Therefore, our test is a conservative test as removing the fixed effects would further increase the estimated positive association.

some participants are confidently wrong, as a positive coefficient implies higher confidence is associated with a higher likelihood of being wrong (since correct answers are coded as 0 in our analyses).

2.1.2 Additional Robustness Checks: Alternative Measure for Prediction Accuracy

We repeat the analysis using a non-binary measure of prediction accuracy (continuous specification of the dependent variable based on Question 3, e.g., the absolute prediction error in stated probability.

We regress prediction accuracy (i.e. the absolute difference between one's probabilistic estimate and the truth) on meta-prediction accuracy with question and participant fixed effects. Robust standard errors clustered at the participant level are used in all regression results. We find that people with a more accurate probabilistic estimate tend to be more accurate in predicting the average probabilistic estimate of other respondents in the group (b = 0.629, 95% CI = [0.48, 0.77], $p < 0.0001$). One percentage point increase in the accuracy of meta-prediction is associated with 0.629 percentage point increase in the accuracy of meta-prediction.

When regressing prediction accuracy on meta-prediction accuracy with question and participant fixed effects on the seventeen statements with an intuitive answer (i.e. the largest city happens to be the state capital), the regression coefficient is b = 0.576 (95% CI = [0.39, 0.76], $p < 0.0001$). The regression coefficient on the remaining 33 states is b = 0.408 (95% CI = [0.22, 0.59], $p < 0.0001$). As a comparison, when we change the independent variable to people's confidence (calculated as one's stated probability of their binary response being correct), the regression coefficients for the statements with intuitive and counterintuitive outcomes are b = -1.01 (95% CI = [-1.14, -0.87], $p < 0.0001$) and b = -0.054 (95% CI = [-0.24, 0.14], p = 0.59), respectively.

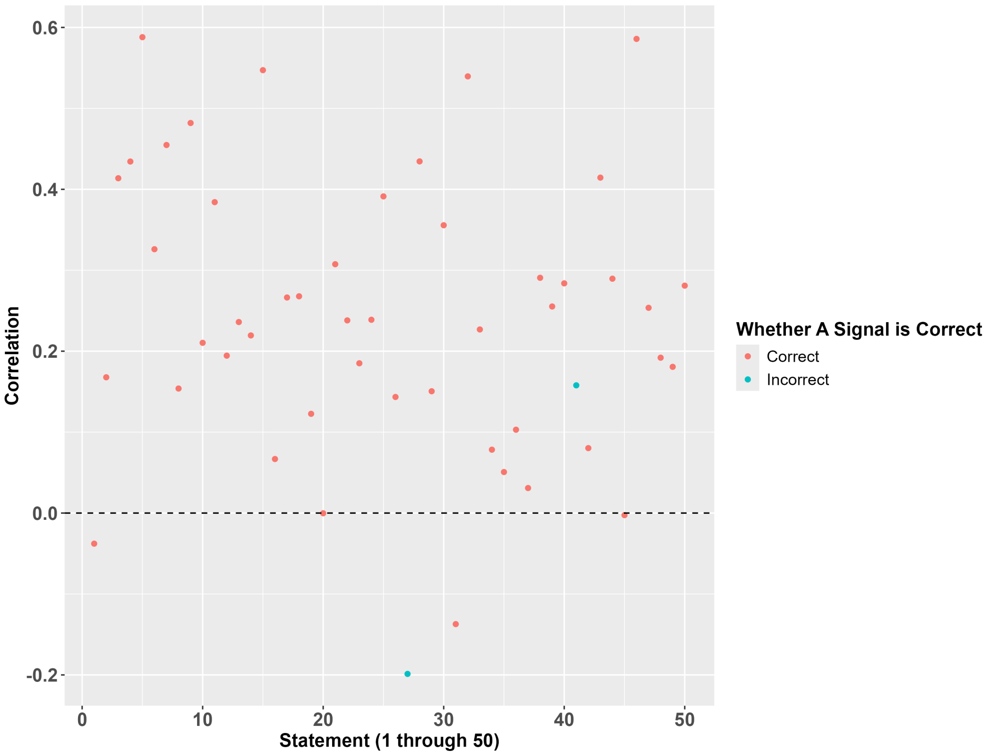

Figure 1: *The x-axis corresponds to Statement 1 through Statement 50 (the exact statement is prepared by Wilkening et al. and included a separate pdf in the materials.) The y-axis is the correlation between prediction accuracy and meta-prediction accuracy for each statement. The color indicates whether our aggregation rule based on meta-prediction accuracy has a correct outcome.*

Furthermore, we also calculate Pearson's correlation between prediction accuracy (i.e. the absolute difference between one's probabilistic estimate and the truth) and meta-prediction accuracy for each statement. We find a positive correlation in 45 out of 50 statements (see Figure 1).

2.2 Context 2: Questions with Varying Difficulty Levels

2.2.1 Context 2 Results: Association between Prediction Accuracy and Meta-Prediction Accuracy

Same as in Context 1, we regress prediction accuracy on meta-prediction accuracy with question and participant fixed effects within each difficulty level, and robust standard errors clustered at the participant level. We find a positive association between prediction accuracy and meta-prediction accuracy for all difficulty levels (see Table 1 below for regression results)

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

Interestingly, as the difficulty level increases, the regression coefficient b becomes smaller. To confirm this analytically, we regress prediction accuracy on meta-prediction accuracy, difficulty level, and their interaction, with question and participant fixed effects, and robust standard errors clustered at the participant level. The coefficient on the interaction term is significantly negative ($b = -0.087$, $t = -7.4$, 95% CI = [-0.11, -0.06], $p < 0.0001$), which means the association between prediction accuracy and meta-prediction accuracy is weakened as questions become more difficult. One potential explanation is that as the difficulty of problems increases, it becomes more difficult for the "experts" (e.g., those who have an accurate binary prediction) to correctly estimate the crowd composition relative to the laypeople (e.g., those whose binary prediction is inaccurate), thereby reducing the accuracy of their meta-predictions.

Finally, unlike in the previous context in which confidence does not predict accuracy when the truth is counter-intuitive, we find a significantly negative association between confidence and prediction accuracy at each difficulty level (see the regression coefficients with confidence as the independent variable below). In order to compare whether confidence or meta-prediction accuracy better associates with prediction accuracy overall, we perform two regression models. The first model regresses prediction accuracy on standardized meta-prediction accuracy, with question and participant fixed effects, and robust standard errors clustered at the participant level – pooling all five hundred questions. The coefficient on meta-prediction accuracy is 0.101 ($t = 19.05$, $p < 0.0001$), and the model's adjusted R^2 and AIC value are 0.255 and 46,655, respectively. The second model is the same regression except with the independent variable being standardized confidence. The coefficient on confidence is -0.070 ($t = -22.5$, $p < 0.0001$), and the model's adjusted R^2 and AIC value are 0.235 and 47,872. The magnitude of the coefficient and the adjusted R^2 in the standardized meta-prediction accuracy model are larger, the its AIC value smaller. These results suggest that standardized meta-prediction accuracy better predicts accuracy than standardized confidence.

2.2.2 Additional Robustness Checks: Alternative Measure for Prediction Accuracy

Difficulty	Coefficient	t-stat	p-value	95% CI
1	0.80	22.6	< 0.0001	[0.73, 0.87]
2	0.676	18.8	< 0.0001	[0.61, 0.75]
3	0.539	13.7	< 0.0001	[0.46, 0.62]
4	0.392	10.8	< 0.0001	[0.32, 0.46]
5	0.461	12.6	< 0.0001	[0.39, 0.53]

The relationship is robust when using the alternative continuous specification for our dependent variable. We regress prediction accuracy (i.e. the absolute difference between one's probabilistic estimate and the truth) on meta-prediction accuracy with question and participant fixed effects, with robust standard errors clustered at the participant level. The positive association is robust using the alternative measurement of prediction accuracy (see Table 2.2.2). We also calculate Pearson's correlation between prediction accuracy (i.e. the absolute difference between one's probabilistic estimate and the truth) and meta-prediction accuracy for each statement. We find a positive correlation in 397 out of 500 statements (Difficulty 1: 94; Difficulty 2: 83; Difficulty 3: 78, Difficulty 4: 70, Difficulty 5: 72). As shown in Figure 2, most instances where a our aggregation algorithm based on meta-prediction accuracy produces an incorrect output are associated with a negative correlation between prediction accuracy and meta-prediction accuracy.

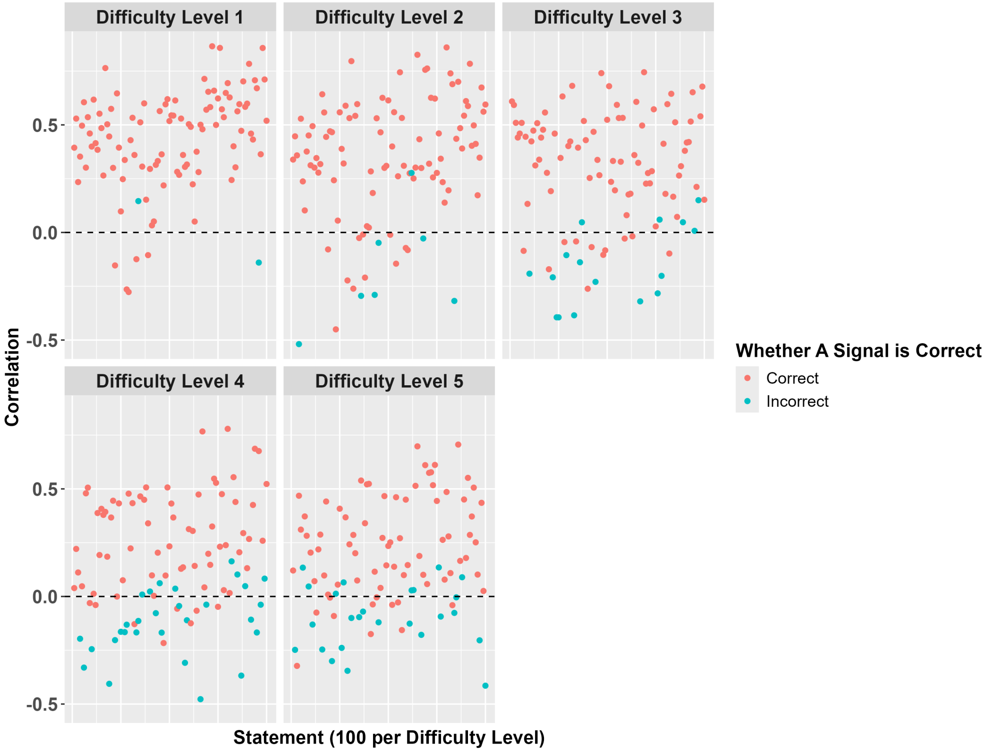

Figure 2: *The x-axis corresponds to Statements 1 through Statement 500 with 100 statements in each difficulty level (the exact statement is prepared by Wilkening et al. and included a separate pdf in the materials.) The y-axis is the correlation between prediction accuracy and meta-prediction accuracy for each statement. The color indicates whether our aggregation rule based on meta-prediction accuracy has a correct outcome.*

2.2.3 Confidence-Accuracy Correlation

For each difficulty level, we regress prediction accuracy (such that an incorrect vote is coded as 1 and a correct vote as 0) on confidence with question and participant fixed effects, with robust standard errors clustered at the participant level. Results are summarized in Table 2-SI.

For each difficulty level, we regress prediction accuracy (i.e. the absolute difference between one's probabilistic estimate and the truth) on confidence with question and participant fixed effects, with robust standard errors clustered at the participant level. Results are summarized in Table 2.

Difficulty Level	Regression Coefficient (b)	t-stat	95% CI of b	p-value
1	-0.54	-13.9	[-0.61, -0.46]	< 0.0001
2	-0.44	-12.1	[-0.51, -0.37]	< 0.0001
3	-0.36	-9.2	[-0.44, -0.28]	< 0.0001
4	-0.33	-9.1	[-0.40, -0.26]	< 0.0001
5	-0.25	-6.4	[-0.32, -0.17]	< 0.0001

Table 2: Regression coefficient (b) characterizes the association between confidence and prediction accuracy at each difficulty level.

Next we conduct a robustness check using the alternative measure for prediction accuracy. For each difficulty level, we regress prediction accuracy (i.e. the absolute difference between one's probabilistic estimate and the truth) on confidence with question and participant fixed effects, with robust standard errors clustered at the participant level. Results are summarized in Table 3. Overall, confidence predicts accuracy in this context. Nevertheless, as shown in the main text, meta-prediction accuracy better predicts accuracy than confidence.

Difficulty Level	**Regression Coefficient (b)**	**t-stat**	**95% CI of b**	**p-value**
1	-0.81	-31.9	[-0.86, -0.76]	< 0.0001
2	-0.59	-24.8	[-0.63, -0.54]	< 0.0001
3	-0.49	-19.9	[-0.54, -0.44]	< 0.0001
4	-0.373	-15.5	[-0.42, -0.33]	< 0.0001
5	-0.367	-15.0	[-0.41, -0.32]	< 0.0001

Table 3: Regression coefficient (b) characterizes the association between confidence and prediction accuracy measured in probability at each difficulty level.

3 Accuracy of Different Judgment Aggregation Methods

After demonstrating the positive association between prediction accuracy and meta-prediction accuracy, we hereby examine the accuracy of different judgment aggregation methods using data provided by Wilkening et al. (2022), which we described in the previous section.

3.1 Performance Evaluation in the 50-state Dataset

For our proposed algorithm based on meta-prediction accuracy, we calculate the average meta-prediction (average estimate of Question 4) by those who responded "true" and those who responded "false" separately. Then we compare the average meta-prediction of each response category against the "meta-prediction truth." In 48 out of 50 cases, participants in the correct response category predicted the average estimate of the crowd more accurately than those in the incorrect category. Thus, our method was inaccurate in only 2 out of 50 statements. We compare the aggregation performance of our method against various baselines in Table 4: the majority rule, the average probabilistic estimate (i.e., using 50% as a threshold for "true" versus "false"), the group-confidence-based method (i.e., determining the aggregate output based on which response category contains agents with higher average confidence in their corresponding category), the minimal pivot method, the Surprisingly Popular (SP) algorithm, and the state-of-the-art knowledge-weighted method.

Question Type	**Aggregation Method**							
	Majority	Average	Confidence	Minimal Pivot	Surprisingly Popular	Knowledge weight	Knowledge weight (outlier robust)	Meta-prediction 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 Inaccuracies (50 Statements)	24 (48%)	23 (46%)	11 (22%)	5 (10%)	7 (14%)	3 (6%)	15 (30%)	2 (4%)

Table 4: *Number of inaccurate answers computed using each aggregation method among statements where the truth is intuitive (17 statements) versus counter-intuitive (33 statements).*

When our algorithm identifies the accurate response category, aggregating over the probabilities within the accurate category must outperform the average probability of the entire crowd which includes probabilistic estimates from the inaccurate category. For example, suppose a statement is true, the average probability among those whose probabilistic prediction is greater than 50%, which is what our algorithm does, must be closer to the ground truth (100%) than the average probability of the entire crowd which includes probabilities below 50%. Indeed, as we can see from Table 5, our method is the best performing method with lowest RMSE.[3]

[3] The omitted methods can only produce categorical output.

Question Type	**Aggregation Method**					
	Average	Confidence	Minimal Pivot	Knowledge weight	Knowledge weight (outlier robust)	Meta-prediction 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 5: *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 Performance Evaluation in the 500-Science-Statement Dataset

We computed the aggregate output of our method based on meta-prediction accuracy and compared it to existing benchmarks. As summarized in Table 7, our method outperforms classic benchmarks across difficulty levels, and is overall on par with advanced methods (i.e. our method performs the best at difficulty levels 1 and 2, but is very close to but not the best at difficulty levels 3 through 5)(see Tables 1 and 7, respectively, for binary and RMSE comparisons).

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 6: *Number of inaccurate answers computed using each aggregation method at each difficulty level. There are 100 questions in total for each difficulty level.*

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 7: *Root Mean Squared Error (RMSE) computed using each aggregation method at each difficulty level. There are 100 questions in total for each difficulty level.*

4 Additional Analysis of the Social Influence Experiment with Binary Signals

4.1 Controlling for the Absolute Error of Initial Probabilistic Estimate

As a robustness check, we regress improvement on conditions (baseline = meta-prediction accuracy influence, factor 1 = majority influence, factor 2 = confidence-based influence, factor 3 = coin flip) with statement fixed effects, adding the absolute error of participants' initial probabilistic estimate as a control (robust standard errors clustered at the participant level). The coefficient on initial absolute error is significantly positive ($b = 0.316, 95\% CI = [0.29, 0.34], p < 0.0001$), which means people who had larger initial error tend to have larger improvement after seeing social influence. All main coefficients are significantly negative as in the main analysis ($b1 = -7.27, 95\% CI = [-9.3, -5.2], p < 0.0001; b2 = -2.53, 95\% CI = [-4.8, -0.3], p = 0.027; b3 = -3.11, 95\% CI = [-5.5, -0.69], p = 0.012$).

4.2 Improvement Among Those Who Initially Have the Correct vs Incorrect Binary Classification

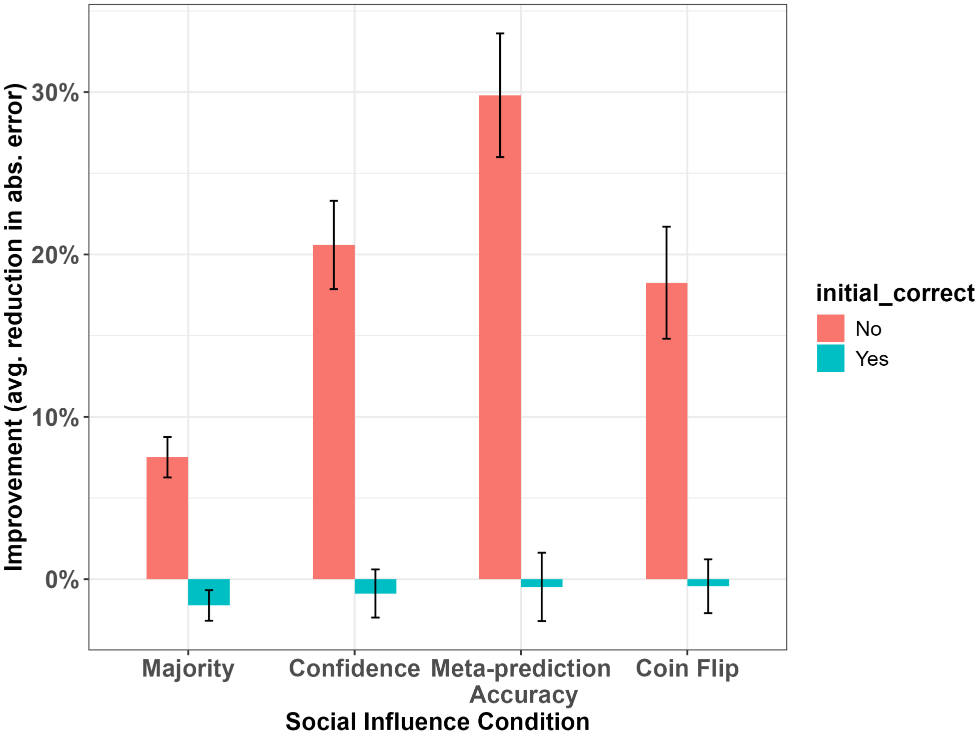

Figure 3: *The x-axis is the four types of social influence. The y-axis is the average improvement in probability estimates.*

As demonstrated in Figure 3, improvement from social influence is mainly driven by people with initially incorrect estimates (e.g. initial probability $< 50\%$ for a true statement or initial probability $> 50\%$ for a false statement). For those whose initial estimate is on the correct side, influence based on meta-prediction accuracy has little to harm on average (t-test, comparing improvement of those with initially correct estimates in the meta-prediction accuracy influence condition to 0; $-0.474\%, t(6719) = -1.70, 95\% CI = [-1.02\%, 0.07\%], p = 0.088$).

4.3 When Improvement Is Measured with Binary Outcome

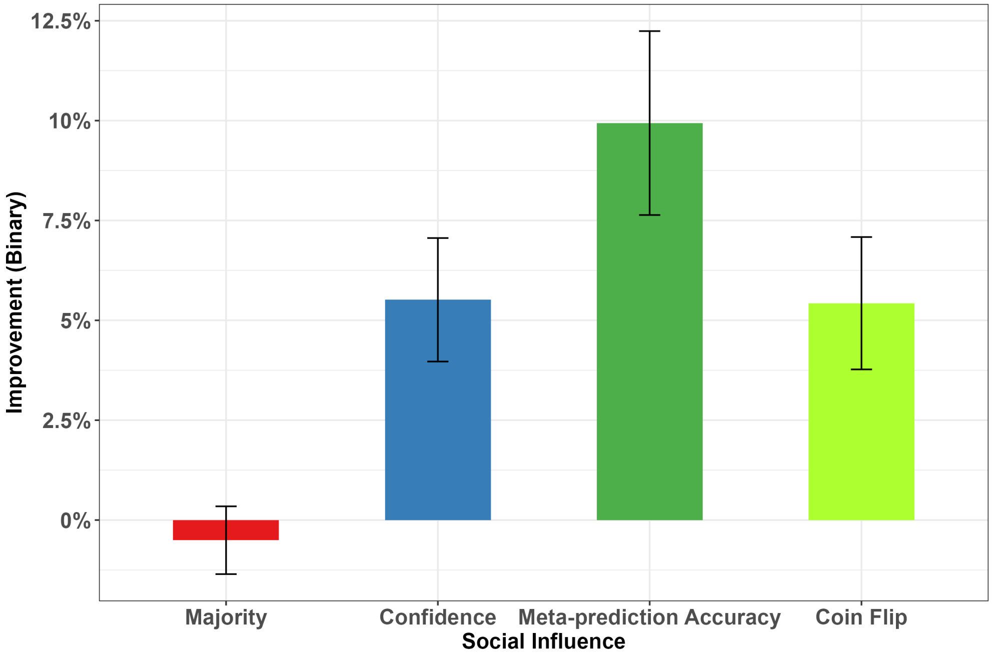

Figure 4: *The x-axis is the four types of social influence. The y-axis represents improvement measured in terms of binary outcomes.*

Instead of calculating improvement using the continuous probabilistic scale, we define a probabilistic estimate as having an accurate binary outcome if the stated probability is less than 50% for a true statement or greater than 50% for a false statement. Then we calculate the DV as the absolute error of prior binary judgment minus the absolute error of posterior binary estimate. A positive (negative) value means the posterior binary judgment becomes accurate (inaccurate) after seeing social influence when the prior binary judgment is wrong (right); 0 means the binary judgment remains the same after seeing social influence.

Combining the fifty statements, we regress the binary improvement on conditions (baseline = meta-prediction accuracy influence, factor 1 = majority influence, factor 2 = confidence-based influence, factor 3 = coin flip) with statement fixed effects (robust standard errors clustered at the participant level). All coefficients are significantly negative ($b1 = -0.104, 95\%CI = [-0.13, -0.08], p < 0.0001; b2 = -0.044, 95\%CI = [-0.072, -0.017], p = 0.002; b3 = -0.045, 95\%CI = [-0.073, -0.017], p = 0.002$). Results are robust when we add the absolute error of participants' initial probabilistic estimate as a control ($b1 = -0.095, 95\%CI = [-0.12, -0.07], p < 0.0001; b2 = -0.028, 95\%CI = [-0.053, -0.004], p = 0.025; b3 = -0.03, 95\%CI = [-0.055, -0.003], p = 0.028$).

4.4 Amount of Belief-updating in the Direction of the Signals

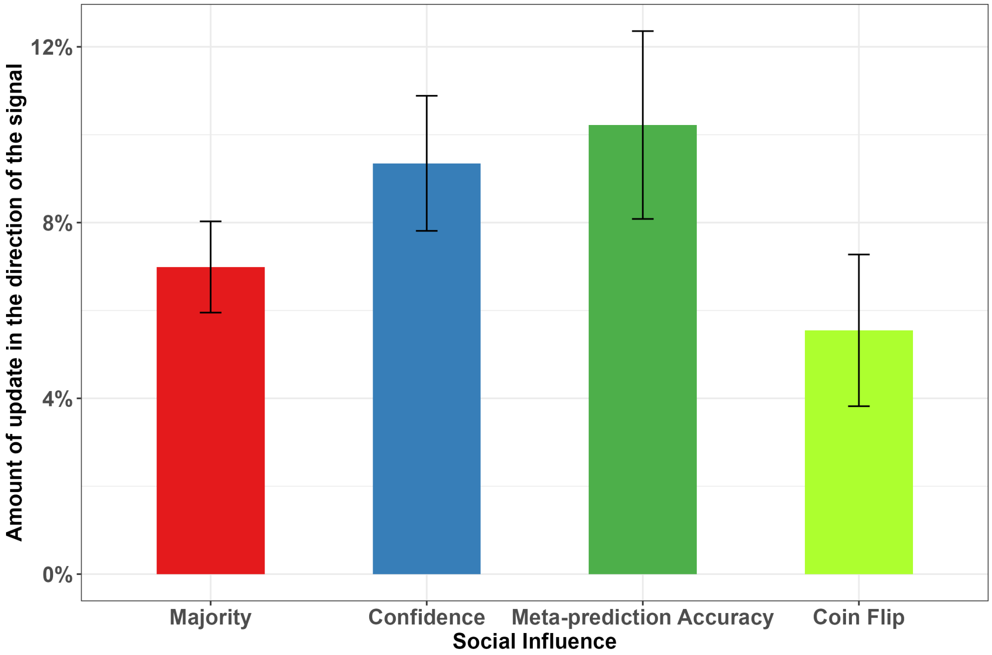

Figure 5: *The x-axis is the four types of social influence. The y-axis is how much people update in the direction of their given signal on average across all statements.*

Next we compare the overall belief-updating across conditions to gain a better understanding of how "influential" different signals are (see Figure 5). For those who receive "True" ("False") as a signal for a statement, we calculate the amount of updating as the posterior (prior) probability of a statement being true minus the prior (posterior) probability of a statement being true. Thus, a positive (negative) value means updating in the (opposite) direction of the signal. This is different from the analysis in the main text in which we examine the extent to which people update toward the truth: improvement could possibly come from people updating in the opposite direction of the signal when the signal is wrong, but updating in the opposite direction also implies people do not trust the signal.

Combining the fifty statements, we regress the amount of belief-updating on conditions (baseline = influence based on meta-prediction accuracy, factor 1 = majority influence, factor 2 = confidence-based influence, factor 3 = coin flip) with statement fixed effects, and robust standard errors clustered at the participant level. The negative coefficients suggest influence based on meta-predictoin accuracy indeed served as influential signals ($b1 = -3.23, 95\%CI = [-5.6, -0.86], p = 0.008; b2 = -0.87, 95\%CI = [-3.5, 1.76], p = 0.52; b3 = -4.67, 95\%CI = [-3.34, -7.42], p = 0.0008$). Results are robust when we add the absolute error of participants' initial probabilistic estimate as a control ($b1 = -2.58, 95\%CI = [-4.70, -2.12], p < 0.017; b2 = 0.23, 95\%CI = [-2.12, 2.59], p =$

$0.85; b3 = -3.58, 95\%CI = [-6.1, -1.04], p = 0.0057$). In particular, because influence based on meta-prediction accuracy and the "coin-flip influence" contain the same signals for the fifty statements, the results indeed indicate that people update their beliefs less if the social influence comes from an uninformative procedure.

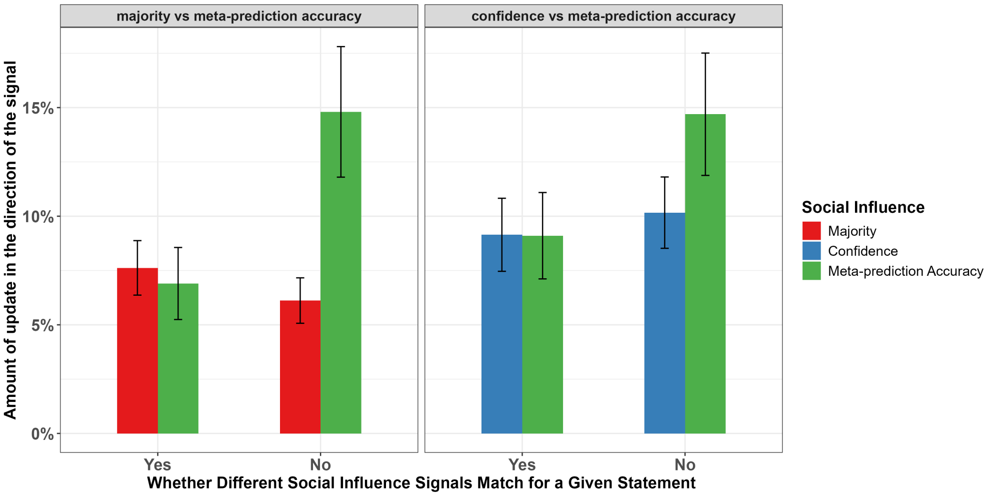

Figure 6: *The y-axis is the average amount of belief-updating in the direction of the signal. The x-axis is whether we are comparing statements in which different methods have the same signal. The left panel compares majority and meta-prediction accuracy influence in the 29 statements where majority and our method based on meta-prediction accuracy produce the same signal and the 21 statements where their signals differ. The right panel compares confidence and meta-prediction accuracy influence in the 40 statements where they produce the same signal and the 10 statements where the signals differ.*

Next we examine people's belief-updating when different methods produce the same versus different signals (see Figure 6). Using the regression introduced in the main text among the 29 statements where majority and meta-prediction accuracy influence have the same signal, the two types of influence induce a similar amount of belief-updating toward the signal ($b = 0.72$, 95% CI = [-1.55, 2.99], $p = 0.53$). Participants update their beliefs substantially more under meta-prediction accuracy influence in the remaining 21 statements where its signals differ from majority influence ($b = -8.69$, 95% CI = [-11.9, -5.5], $p < 0.0001$). Among the 40 statements where confidence and social influence based on meta-prediction accuracy produce the same signal, the two types of social influence induce similar belief-updating ($b = 0.042$, 95% CI = [-2.64, 2.72], $p = 0.98$). Among the remaining 10 statements where the two types of social influence have disagreeing signals, social influence based on meta-prediction accuracy induces larger updates toward the signal than confidence influence ($b = -4.53$, 95% CI = [-7.85, -1.2], $p = 0.008$). Finally, because influence based on meta-prediction accuracy and the "coin-flip influence" contain the same signals for all fifty statements (by design), the results indeed indicate that people update their beliefs less if the social influence comes from an uninformative procedure ($b = -4.67$, 95% CI = [-3.34, -7.42], $p = 0.0008$).

5 Study 2: The Social Influence Experiment with Probabilistic Signals

We hereby provide the full details of the social influence experiment with probabilistic signals in the main text.

5.1 Methods

Full experimental materials including the pre-registration are available on our OSF site. We employ two-tailed tests and identify as "post-hoc" any analyses we did not pre-register.

5.1.1 Participants

A total of 1001 participants from Prolific[4] completed our study. The median time to complete the survey was 16.5 minutes. The average total earnings per participant was $3.60 fixed payment plus an average of $0.25 additional performance-based bonus. 80 participants earned the highest possible reward of $1 for their accuracy.

5.1.2 Procedure

The stimuli and study procedure are exactly the same as in Study 1 except for the social influence signals. The exact social influence messages provided to the participants are included below:

1. In the **Average** condition, participants read: "*On average, the previous group of participants believed that there was a X% probability the statement is true.*"

2. In the **Majority+Average** condition, participants read: "*In the previous group of participants, more people rated the statement as more likely to be "True/False". On average, the previous participants believed that there was a X% probability the statement is true.*"

3. In the **Confidence** condition, participants read: "*In the previous group of participants, the people who rated the statement as more likely to be "True/False" reported more confidence in their answer than those who rated the statement as more likely to be "False/True". On average, those who rated the statement as more likely to be "True/False" believed that there was a Y% probability the statement is true.*".

4. In the **Meta-prediction accuracy** condition, participants read: "*In the previous group of participants, the people who rated the statement as more likely to be "True/False" could more accurately predict the average of other people's estimates than those who rated the statement as more likely to be "False/True". On average, those who rated the statement as more likely to be "True/False" believed there was a Z% probability the statement is true.*"

[4] We recruited samples quota-matched to the U.S. population, and did not ask for demographic information in the survey.

We also have a separate condition that aims to mimic the “coin flips” influence condition introduced in the previous study with binary signals. In particular, participants in this “random number influence” condition see the following message: *“The random number generator returned the number Z%, which randomly identifies the statement as ”True/False” and that there is a Z% probability the statement is true.”* In this condition, unbeknownst to the participants, the binary and the probabilistic outcomes from the random number generator are matched to the outcome of the meta-prediction accuracy influence condition. In particular, it is the choice of the group with the highest meta-prediction accuracy.

5.2 Results: Improvement in Prediction Accuracy Under Social Influence

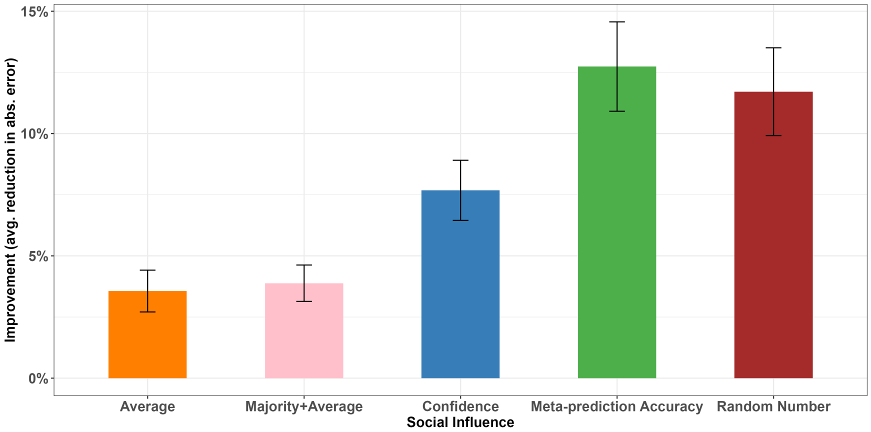

Figure 7: *The x-axis is the five types of social influence conditions. The y-axis is the average improvement in participants' probability estimates from before versus after they received the social influence information. The bars indicate 95% confidence intervals based on robust standard errors.*

First, as visualized in Figure 7, social influence based on meta-prediction accuracy resulted in an average improvement of 12.7 percentage points. We regress the improvement on the assigned conditions with the meta-prediction accuracy influence as the baseline (factor 1 = average influence, factor 2 = majority + average influence, factor 3 = confidence-based influence, factor 4 = random number influence), controlling for the absolute error of participants' initial probabilistic estimate, along with statement fixed effects, and robust standard errors clustered at the participant level. Most regression coefficients are significantly negative ($b_{average} = -9.64, 95\% CI = [-11.48, -7.79]$; $b_{majority+average} = -9.63, 95\% CI = [-11.41, -7.86]$, $b_{confidence} = -4.61, 95\% CI = [-6.43, -2.78]$, all $ps < 0.0001$; $b_{random} = -0.84, 95\% CI = [-0.75, -3.01], p = 0.451$), which means meta-prediction accuracy influence on average resulted in a larger improvement compared to the other types of social influence except for the random influence condition. Nevertheless, we do not think that a null comparison against the random influence means people treat our meta-prediction accuracy influence as non-informative anchoring. It is more likely that – as the study progresses – some participants notice that the probabilistic signals generated by what the researchers claim as the “random number generator” are likely not random. In other words, the result likely indicates that our experimental manipulation failed here because participants do not believe that the "random signals" are random. In the previous study, this is perhaps a less concerning problem because the communicated signals were binary.

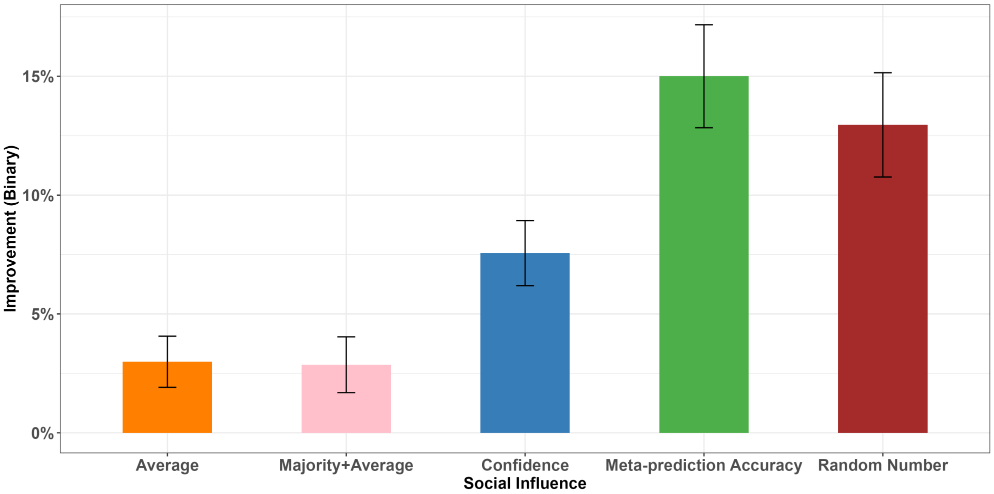

Figure 8: *The x-axis is the five types of social influence. The y-axis represents improvement measured in terms of binary outcomes.*

The results, visualized in Figure 8, are robust when we replace the dependent variable in the previous regression with whether a prediction changes from "incorrect" to "correct" ($b_{average} = -0.126, 95\%CI = [-0.148, -0.103]$; $b_{majority+average} = -0.131, 95\%CI = [-0.154, -0.109]$, $b_{confidence} = -0.069, 95\%CI = [-0.090, -0.047]$, all $ps < 0.0001$; $b_{random} = -0.018, 95\%CI = [-0.044, 0.008], p = 0.171$).

5.3 Improvement Among Those Who Initially Have the Correct vs Incorrect Binary Classification

As shown in Figure 9, when probabilistic signals are communicated, most improvement comes from responses that were initially wrong (e.g., these responses had more room for improvement). This is consistent with our findings in the previous social influence experiment in which we communicated binary signals. One thing worth noting is that the negative social influence effect on those with correct initial responses is mitigated when we communicate probabilistic, as opposed to binary signals (t-test, comparing improvement of those with initially correct estimates in the meta-prediction accuracy influence condition to 0; $1.50\%, t(6761) = 6, 95\%CI = [1.04\%, 1.95\%], p < 0.0001$). This phenomenon – mostly driven by initial estimates that were close to the 50% cutoff – has been discussed in detail in Becker et al. (2022).

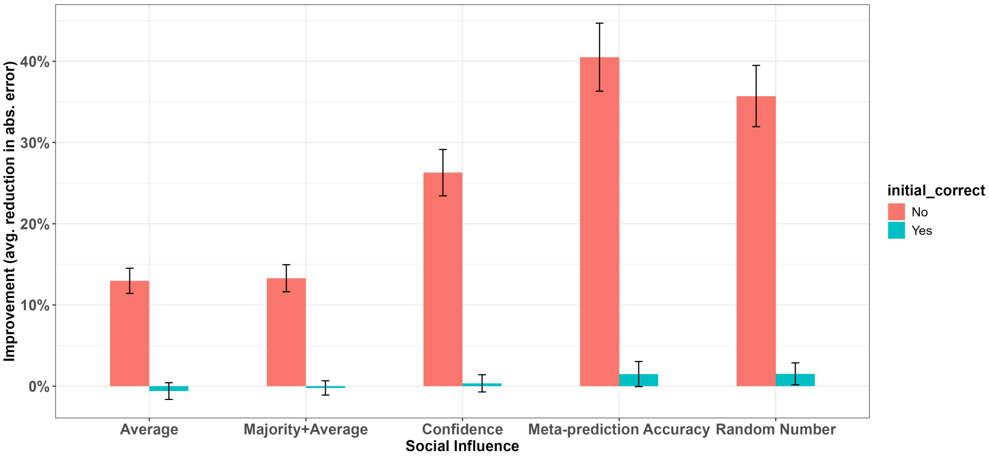

Figure 9: *The x-axis is the five types of social influence. The y-axis represents improvement measured in terms of probabilistic outcomes. The colors indicate responses that were initially correct (evaluated in a binary fashion) versus wrong.*

5.4 Amount of Belief-updating in the Direction of the Signals

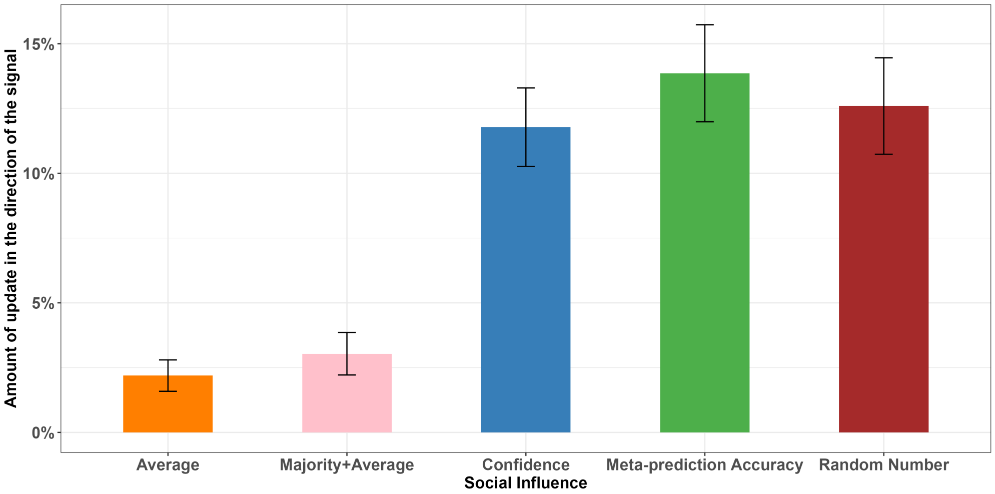

Figure 10: *The x-axis is the social influence conditions. The y-axis is the average amount of updating in probability estimates. The bars indicate 95% confidence intervals based on cluster robust standard errors.*

Figure 11 depicts the amount of belief-updating in the direction of the signal in each of the social influence conditions. Using the same regression as in Section 5.2 with the amount of belief-updating being the dependent variable, we find that social influence based on meta-prediction accuracy induced a larger amount of belief-updating than the "Average" and the "Majority+Average" influence ($b_{average} = -12.01, 95\%CI = [-13.80, -10.21]$, $p < 0.0001$; $b_{majority+average} = -11.39, 95\%CI = [-13.25, -9.54]$, $p < 0.0001$). This is because probabilistic signals based on meta-prediction accuracy are usually more extreme than the other two types of social influence, as the latter two types of influence incorporate probabilistic estimates from both the "True" group and the "False" group. In addition, social influence based on meta-prediction accuracy resulted in a similar degree of belief-updating compared to the confidence-based influence and the "random number" influence ($b_{confidence} = -1.75, 95\%CI = [-3.781, 0.288]$, $p = 0.093$; $b_{random} = -1.12, 95\%CI = [-3.42, 1.18], p = 0.338$).

As one may observe from Figure 11, among the 40 statements in which confidence and meta-prediction accuracy have the same signals, the two types of social influence yield a similar average amount of belief-updating ($b = -1.02$, 95% CI = [-3.46, 1.43], $p = 0.416$).[5] This means that people's belief-updating under the meta-prediction accuracy influence is the same as that under the confidence-based influence, which implies that people trust the signals based on meta-prediction accuracy to the extent they trust the well-established signals based on confidence. However, when the signals disagree, social influence based on meta-prediction accuracy yields a substantially larger amount of belief-updating than confidence-based influence ($b = -5.86$, 95% CI = [-8.89, -2.84],

[5] The regression is the same as in the main text, except the dependent variable is the amount of belief-updating.

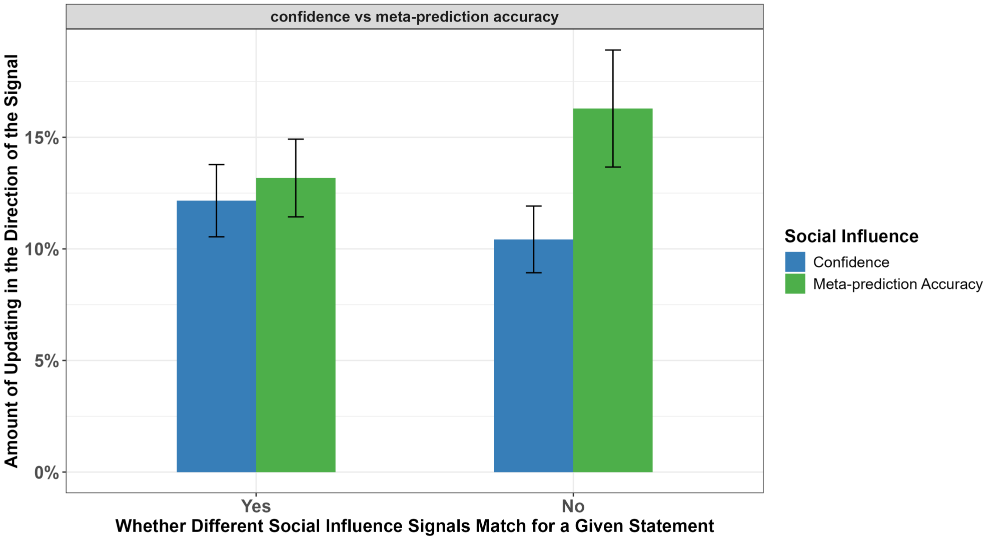

Figure 11: *The x-axis is whether we are comparing statements in which confidence and meta-prediction accuracy have the same signal. The y-axis is the average amount of updating in probability estimates. The bars indicate 95% confidence intervals based on cluster robust standard errors.*

$p = 0.0001$). This result is perhaps a reflection of the difference in signals as opposed to trust.

6 Palley and Soll (2019): Questions with Outcomes that are More Aleatory

In their Study 4, Palley and Soll (2019) incentivized participants (N ranges between 48 and 165) to guess the probability that each team would win across 120 different games in the early rounds of the 2014, 2015, and 2016 NCAA Division I Men's Basketball Tournaments (also known as "March Madness").[6] For each game, they asked their participants to:

- Question 1: predict the winner of the game;
- Question 2: assign a probability that their selected team would win;
- Question 3: estimate the average probability of winning that other participants would assign to their selected team.

While the knowledge-based questions in Wilkening et al. (2022) represent more epistemic contexts, the basketball game outcomes in this context are more aleatory, meaning it is inherently difficult (if not impossible) to perfectly predict them. For example, even if an oracle knows Team A has a 70% chance of winning against Team B, this implies that in 100 games played in parallel universes, Team A would win approximately 70 times and lose 30 times. Therefore, the outcome we observe is analogous to the result of a biased coin flip with a 70% chance of heads, and we

[6] Studies 1 and 2 in Palley and Soll (2019) are not relevant in the context of our paper.

should expect 30% of guesses to be wrong. There is no deterministically correct answer for these problems, *a priori*.

To obtain the underlying true probability, the authors converted the betting odds obtained from various betting websites to an average probability of a team winning. Across the 120 games, the market predicted the "overdog team" (e.g., the team predicted to have more than 50% chance of winning) would win with 74.2% probability on average, and the "overdog team" won in 77.5% of the match-ups (93 out of 120 games), which suggests the market is a good proxy of the underlying probability.

6.1 Association between Prediction Accuracy and Meta-Prediction Accuracy

First, aligned with our analysis in the previous two contexts, we regress prediction accuracy, measured against the actual game outcome, on meta-prediction accuracy with question and participant fixed effects, and robust standard errors clustered at the participant level. We find a positive association between prediction accuracy and meta-prediction accuracy in this aleatory context with no deterministically correct answers *a priori* ($b = 0.216$, $t = 3.85$, 95% CI = [0.11, 0.33], $p = 0.0001$). The coefficient when we use standardized meta-prediction accuracy as the independent variable is 0.0266 ($t = 3.85$, 95%CI = [0.013, 0.040], $p = 0.0001$). This association remains robust when we specify our dependent variable in a continuous space[7] using market probability as the ground truth ($b = 0.386$, $t = 15.1$, 95% CI = [0.34, 0.44], $p < 0.0001$). This suggests that one percentage point increase in the accuracy of meta-prediction is associated with 0.386 percentage point increase in the accuracy of one's prediction of market probability.

We also regress prediction accuracy based on actual game outcomes on standardized confidence, with question and participant fixed effects, and robust standard errors clustered at the participant level. The coefficient is -0.053 ($t = -8.6$, 95% CI = [-0.065, -0.041], $p < 0.0001$). Comparing the coefficients, we find that confidence better associates with prediction accuracy than does meta-prediction accuracy in this context. This is probably because match-ups during "March Madness" are based on seeding, based on which participants can better calibrate their confidence.

6.2 Performance of Different Aggregation Methods

We hereby examine the performance of different aggregation methods in this context. We will compare the aggregate outcomes with both the actual game outcomes and the betting market prediction (e.g. assuming the betting market's prediction as the ground truth.)

Ground Truth	Aggregation Method						
	Majority	Average	Confidence	Minimal Pivot	Knowledge weight	Knowledge weight (outlier robust)	Meta-accuracy
Actual	30	30	31	30	30	30	29
Market	15	13	16	13	13	13	14

Table 8: *Number of inaccurate answers computed using each aggregation method. The actual game outcome is binary. The "Market" ground truth is the binary outcome of market prediction (e.g. a market prediction greater than 50% is "Team A wins", and a probability below 50% is "Team B wins".)*

6.2.1 Performance Evaluation in terms of Binary Outcomes

As shown in Table 8, the performance across different methods is similar. This is presumably because (1) conventional aggregation methods are already fairly accurate – leaving less room for improvement; and (2) the context is fairly aleatory, which means the outcomes are less predictable (and one cannot predict "randomness.").

6.2.2 Performance Evaluation in terms of Probabilistic Outcomes

Ground Truth	Aggregation Method					
	Average	Confidence	Minimal Pivot	Knowledge weight	Knowledge weight (outlier robust)	Meta-accuracy
Actual	0.434	0.424	0.420	0.431	0.429	0.431
Market Probability	0.150	0.145	0.140	0.142	0.144	0.137

Table 9: *Root Mean Squared Error (RMSE) computed using each aggregation method. The actual game outcome is binary. The "Market" ground truth is the market's probabilistic prediction.*

Overall, the performance across different methods in this aleatory context is similar.

[7] Prediction accuracy is measured by the absolute difference between one's stated probability of a team winning based on Question 2 and the market probability.