Opinion Averaging versus Argument Exchange

Ulrike Hahn (u.hahn@bbk.ac.uk)
Centre for Cognition, Computation, and Modelling, Birkbeck, University of London
Malet Street, London, WC1E 7HX U.K

Leon Assaad (l.Assaad@campus.lmu.de)
Munich Center for Mathematical Philosophy, LMU Munich, Geschw. Scholl Platz 2, 80539 Munich, Germany

Jason W. Burton (jasonwilliamburton@gmail.com)
Department of Digitalization, Copenhagen Business School, Howitzvej 60, 2000 Frederiksberg, DK
Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, DE

Abstract

Opinion averaging is a common means of judgment aggregation that is employed in the service of crowd wisdom effects. In this paper, we use simulations with agent-based models to highlight contexts in which opinion averaging leads to poor outcomes. Specifically, we illustrate the conditions under which the optimal posterior prescribed by a normative model of Bayesian argument exchange diverges from the mean belief that would be arrived at via simple averaging. The theoretical and practical implications of this are discussed.

Keywords: opinion dynamics; averaging; argument; wisdom of the crowd; agent-based simulation

Introduction

The idea that averaging judgments or opinions is beneficial is both ubiquitous, and it can draw on multiple lines of support. Averaging seems to be a natural strategy for humans both across multiple pieces of evidence or cues (Hogarth & Einhorn, 1992; Hogarth & Karelaia, 2007) and in contexts of social communication (e.g., Jönsson, Hahn, & Olsson, 2015; Becker, Brackbill, & Centola, 2017). This prevalence is matched with convergent indicators on the accuracy benefits of averaging. For one, there are many contexts where a simple averaging strategy well approximates optimal Bayesian inference (see e.g., Juslin, Nilsson, & Winman, 2009); likewise, there are many contexts where simple averaging models outperform more complex models (e.g., Dawes, 1979). The benefits of averaging are also apparent from a sprawling, multi-disciplinary, literature on judgment and/or model aggregation (e.g., Wallsten, Budescu, & Tsao, 1997). Finally, the benefits of averaging are manifest in the demonstration of wisdom of the crowd effects (e.g., Stroop, 1932; Surowiecki, 2004; Becker et al., 2017; Einhorn, Hogarth, & Klempner, 1977), whereby the accuracy of a group mean exceeds that of the individual judgments. Such wisdom of the crowd effects are not a surprising empirical finding. Rather, there is a well-developed, formal, mathematical basis that explains their occurrence, as we detail below.

That mathematical basis also sheds light on a further, related phenomenon: In group contexts involving communication across individuals, it has long been known that communication may, in fact, serve to reduce wisdom of the crowd effects because of the correlation communicative exchange may induce (e.g., Hogarth, 1978). This in turn has prompted an empirical and modelling literature that has sought to determine the implications of this for how best to structure communication and communication networks so as to maximise collective performance (e.g., Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Jönsson et al., 2015; Hahn, Hansen, & Olsson, 2018; Burton, Almaatouq, Rahimian, & Hahn, 2024; Almaatouq et al., 2020).

In this paper, we use an agent-based model to clarify different contexts of opinion/judgment heterogeneity and their implications for the benefits that will accrue through averaging. Through simulation, we illustrate the conditions under which the optimal posterior prescribed by a normative model of Bayesian argument exchange diverges from the mean belief that would be arrived at via simple averaging. This serves to clarify when communication is likely to hurt crowd wisdom, and when not. It serves also to clarify how useful a strategy averaging is across different forms of disagreement. Hence, our results have theoretical implications for the rationality of averaging and practical implications for the design of collective intelligence processes.

The Benefits of Averaging: Formal Background

Multiple formal frameworks elucidate the benefits of averaging and help explain when and why averaging will be beneficial. For example, where a group of estimates is naturally construed as noisy, independent estimates of a true underlying quantity, error cancellation and the central limit theorem (e.g., Surowiecki, 2004) clarify why the average of those estimates will be more accurate.

The Diversity Prediction Theorem (Page, 2006) relates the accuracy of the mean of a group of estimates to the average error of those estimates taken individually and the variance of those estimates. The theorem shows that when accuracy is measured by the squared distance of an estimate to the true value (as with the Brier score, Brier (1950), which is used widely to score the accuracy of probabilistic forecasts) the average individual accuracy is equal to the accuracy of the mean estimate plus the variance. This, in turn, implies the existence of a 'wisdom of the crowd' effect the moment there is non-zero variance in the estimates as:

$$(\bar{x}-\theta)^2 = \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{n} - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \qquad (1)$$

where n is the number of group members, x_i is the estimate of the group member i, $\bar{x}$ is the group's mean estimate, and θ

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

is the true value of whatever is being estimated. Put simply, the theorem means that the squared error of the mean estimate (i.e., the collective error) is equal to the difference between the average individual error squared and the variance of the individuals' judgements. This also makes clear that variance and individual error play an equal role in determining the magnitude of the accuracy boost achieved by averaging.

Finally, Hogarth used a framework drawn from educational testing (Ghiselli, 1964) to clarify performance in group judgment tasks within social psychology (Hogarth, 1978). This analysis considers the correlations between a target value and estimates provided by individual raters, as well as the correlation between target values and the mean ratings. It shows that for k raters x and target values y:

$$\rho_{y\bar{x}} = k^{1/2} * \bar{\rho}_{yx} / [1 + (k-1)\bar{\rho}_{x_i x_j}]^{1/2}$$

(2)

where $\rho_{y\bar{x}}$ is the correlation between the true values and the group's mean estimates, $\bar{\rho}_{yx}$ is the average correlation between the true values and individuals' estimates, and $\bar{\rho}_{x_i x_j}$ is the average correlation between raters' estimates.

In other words, the correlation between the average estimates and the target values is a function of the number of raters, the average individual correlation and the pairwise correlations between the raters themselves.

In their own way, each of the formal results highlight the possible negative impact of communication. Communication undermines the independence of judgments, it is likely to reduce the variance within the group, and it is likely to increase the degree of correlation between raters. From the perspective of error cancellation, this means the central limit theorem no longer applies and convergence toward the true parameter is undercut. In the context of the diversity prediction theorem, reduced variance diminishes the gap between collective and average individual judgments (all other things equal). In the correlational framework of Hogarth/Ghiselli, finally, the mean correlation between raters, together with the mean individual accuracy, determines the limit toward which the correlation between mean (collective) judgment and true values converges as the number of raters increases. This highlights explicitly the cost to the wisdom of crowd effect that might arise from communication. For communication to be beneficial overall, that cost must be outweighed by a sufficient increase in average individual accuracy.

The same fundamental dynamic emerges also from results on judgment aggregation by voting (Ladha, 1992): the increase to individual accuracy must outweigh the costs of the increased dependence if communication is to be a net benefit to collective accuracy (see also, e.g., Hahn, 2022). This has prompted concern about negative impacts of communication for crowd wisdom and experimental investigations probing such negative effects (e.g., Lorenz et al., 2011; Jönsson et al., 2015). The simulations presented in the present work seek to illuminate further the contexts in which both averaging and communication will and will not be beneficial.

Models and Simulations

For our agent-based model we used a recent modelling framework called NormAN—short for 'Normative Argument Exchange across Networks' (Assaad et al., 2023). This framework was created to facilitate agent-based models that involve the exchange of individual arguments across a social network instead of modelling communication simply as opinion averaging (DeGroot, 1974; Lehrer & Wagner, 1981; Hegselmann & Krause, 2002) or as a contagion process (e.g., Centola, 2018). In this it follows on from earlier models of argument exchange such as Mäs and Flache (2013); however, it uses Bayesian agents situated in a ground truth world so that accuracy can be investigated (for discussion, see Assaad et al., 2023).

We modified the publicly available Norman version 1.0 (available here) to include an alternative, opinion averaging procedure based on the Hegselmann-Krause model (Hegselmann & Krause, 2002)[1]. This allows us to compare how opinion averaging would fare for the same initial belief distribution as held by the Bayesian agents.

Evidence, Arguments, and Beliefs in NormAN

In keeping with the NormAN framework, our model involves an underlying 'world' represented by a Bayesian belief network (BN) that is used to specify the true state of the target claim at issue, and to initialise the arguments that are available in principle (Figure 1). To this end, the simulation randomly selects a truth value for the target hypothesis. In other words, it is determined whether in the simulated 'world' of this model run, the target hypothesis is true or false. The value of the hypothesis node in the world BN is then set to that value. This leads to new, revised, (marginal) probabilities for the evidence nodes (the orange nodes in Fig. 1). These probabilities are then used by a random binomial process to determine the state of that evidence node on this particular model run. In other words, the fact that, say, the left-most evidence node (labelled 'one') has a .8 probability of being true given the truth of the hypothesis, means that, on average, 80% of model runs will have this piece of evidence be true in the ground truth world.

Across different model runs, a single underlying 'world' BN will consequently give rise to many different combinations of truth or falsity of the target hypothesis and possible evidence, both for and against. To illustrate further with a simple example. Imagine that the BN in Fig. 1 represents evidence in a criminal trial. The target hypothesis (blue) represents whether Bob committed the crime. The orange evidence nodes represent nine pieces of evidence of varying diagnostic value, such as witness reports or physical evidence that might (or might not) tie Bob to the crime. On a given model run, the value of "Bob committed the crime" is randomly set to true or false, and all possible evidence is randomly generated accordingly: e.g., on this run, witness 1 says he saw Bob in the

[1] Our model code, simulation data, and analysis scripts are publicly available on The Open Science Framework (Link)

vicinity of the crime (i.e., evidence node 1 is set to true). On a different run, however, witness 1 might assert that he did not see Bob at the scene of the crime (i.e., evidence node 1 is set to false), and so on. In short, an underlying BN such as that of Fig. 1 will generate different possible ground truth 'states of the world' across different runs. Crucially, however, the distribution of these will be determined by the structure of the world as set out by the BN. The point of this, in NormAN, is that the nature of the available evidence—that is, how many arguments for and against might, in principle, be found with respect to an empirical claim—is constrained by the nature of the world. It is not arbitrary. In keeping with this, it is less likely (though possible) that there exist more and stronger arguments in favour of a claim that is ultimately false, than in favour of the claim that is ultimately true (for further discussion of this point, see Assaad et al., 2023; Hahn, 2023). We return to the significance of this for our results below.

Having assigned a truth value to both the hypothesis and all possible pieces of evidence at the beginning of the run, those pieces of evidence then form the possible arguments (e.g., 'witness 1 saw Bob at the scene of the crime') that the individual agents in the model can exchange over the course of the simulation.

At the start of a run, agents are randomly assigned individual pieces of evidence. How many such pieces of initial evidence is a free parameter in the model. They then use their own copy of the underlying Bayes' net to update their belief in the target hypothesis in light of the evidence they have received. When communicating with their neighbours—where the probability of communication at each time step is again a free parameter in the model—the agents choose a (single) piece of evidence from their memory according to a communication rule, and communicate that piece of evidence faithfully to their neighbour(s). 'Arguments' in this model are thus communications about evidentially relevant states of the world. Of the pre-configured communication rules in NormAN 1.0, we only use the 'random share' rule in the simulations reported here. As the name suggests, this rule has agents simply randomly choose a piece of evidence from memory to communicate as an argument. That communication, finally, happens across a small-world network (Watts & Strogatz, 1998) in the present simulations. All of this, so far, corresponds to the basic NormAN model as set out in (Assaad et al., 2023).

We added to this setup the possibility of opinion averaging between agents. Specifically, in order to assess the utility of opinion averaging, we clone each agent in the model, providing that clone with the same initial (pre-communication) degree of belief. That degree of belief forms the initial opinion for an independent opinion dynamics process based solely on averaging that unfolds in parallel: while NormAN agents exchange evidence, the clones, at each time step, adopt the average of their own belief and those of their link neighbours. This setup allows us to compare the dynamics of the averaging model with those of (initially perfectly matched) Bayesian agents exchanging evidence, and to examine the accuracy of both vis-à-vis the true state of the world.

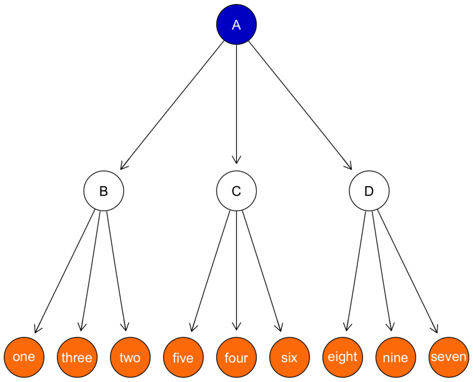


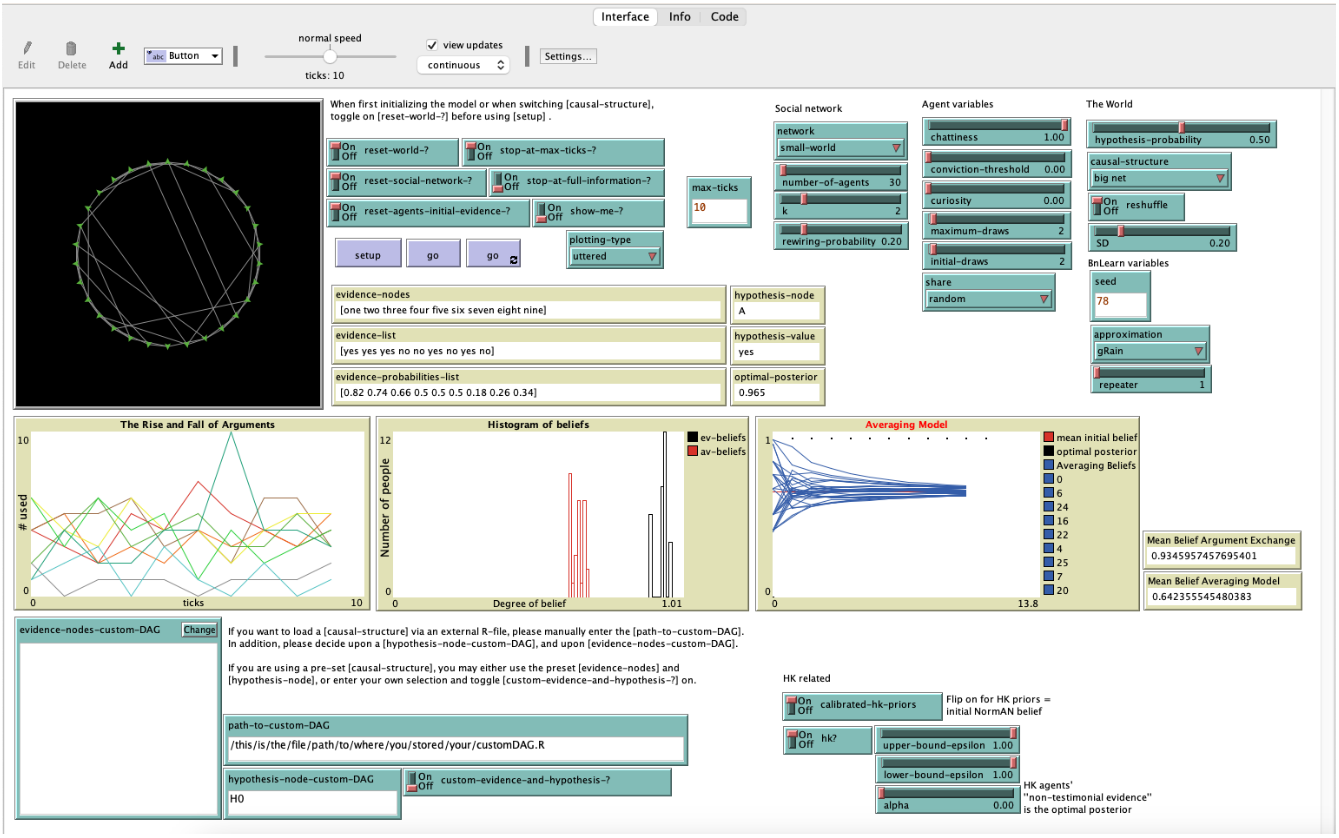
Figure 1: The "big net" world: A Bayesian causal graph used to generate an underlying 'world' for the simulations in NormAN. On a given run, the simulation generates a truth value for a target hypothesis (blue) and truth values for the potential evidence nodes 1 to 9 (orange). These evidence states constitute the full set of potentially available evidence/arguments on that simulation run. Image courtesy of Assaad et al. (2023).

Figure 2 shows a sample run of the model, where there are three main things to note. First, the opinion averaging agents end up a long way from the optimal posterior as determined by the underlying world model: the optimal posterior given (all) evidence in the example is .965, whereas the averaging agents, after 10 steps of exchange, have a mean belief of .64. Second, after those 10 steps the averaging agents also have a very different belief distribution to the argument-exchanging agents who cluster around the optimal posterior with no overlap to the averaging agents (see the output monitor "Histogram of beliefs" in Figure 2), though both populations started with an identical belief distribution.

These divergences occur because exchanging evidence as arguments pulls the Bayesian agents to the optimal posterior, whereas averaging pulls agents to the initial mean belief.

These two very different points of asymptotic convergence are wholly unsurprising as they are intrinsic to the respective models: the optimal posterior is *defined* as the posterior degree of belief (formed on the Bayesian network) given all available evidence, so Bayesian agents who (randomly) exchange will converge to that posterior as their individual knowledge of the evidence set expands. By contrast, averaging models average, so inherently converge toward the population mean, with distortions arising only through differential weighting (whether these be through selective 'trust', network effects, or other mechanisms).

The more interesting point is, consequently, understanding when and why optimal posteriors and initial mean belief will significantly diverge.



When first initializing the model or when switching [causal-structure], toggle on [reset-world-?] before using [setup] .

On Off reset-world-?
On Off stop-at-max-ticks-?
On Off reset-social-network-?
On Off stop-at-full-information-?
On Off reset-agents-initial-evidence-?
On Off show-me-?
max-ticks 10
plotting-type uttered

network small-world
number-of-agents 30
k 2
rewiring-probability 0.20

chattiness 1.00
conviction-threshold 0.00
curiosity 0.00
maximum-draws 2
initial-draws 2
share random

hypothesis-probability 0.50
causal-structure big net
On Off reshuffle
SD 0.20

seed 78
approximation gRain
repeater 1

evidence-nodes [one two three four five six seven eight nine]
hypothesis-node A
evidence-list [yes yes yes no no yes no yes no]
hypothesis-value yes
evidence-probabilities-list [0.82 0.74 0.66 0.5 0.5 0.5 0.18 0.26 0.34]
optimal-posterior 0.965

Mean Belief Argument Exchange 0.9345957457695401
Mean Belief Averaging Model 0.642355545480383

evidence-nodes-custom-DAG

If you want to load a [causal-structure] via an external R-file, please manually enter the [path-to-custom-DAG]. In addition, please decide upon a [hypothesis-node-custom-DAG], and upon [evidence-nodes-custom-DAG].

If you are using a pre-set [causal-structure], you may either use the preset [evidence-nodes] and [hypothesis-node], or enter your own selection and toggle [custom-evidence-and-hypothesis-?] on.

path-to-custom-DAG /this/is/the/file/path/to/where/you/stored/your/customDAG.R
hypothesis-node-custom-DAG H0
On Off custom-evidence-and-hypothesis-?

On Off calibrated-hk-priors
Flip on for HK priors = initial NormAN belief
On Off hk?
upper-bound-epsilon 1.00
lower-bound-epsilon 1.00
alpha 0.00
HK agents' "non-testimonial evidence" is the optimal posterior

Figure 2: The figure shows the NetLogo interface for a sample run of the model. Green boxes show parameter settings in the model. Beige boxes are output monitors that provide information about a particular model run. The six output monitors adjacent to the image of the communication network describe the evidence and hypothesis node in the simulation (see Fig. 1) and their truth values on the given run, as well as the optimal posterior in light of all, in principle, available evidence. The histogram of beliefs below shows the current belief distribution among the regular Bayesian agents. The output monitor to the right with the red header "Averaging Model" shows how the beliefs of the individual agents develop across time. The optimal posterior (black dashed line) and the mean belief of the agents (red line) are added for comparison.

Divergence Between Optimal Posterior and Initial Mean Belief

To give an indication of the prevalence of substantial divergence, we used one of the pre-defined 'world models' contained in NormAN version 1.0, the "big net" world model of Figure 1.

The first three of its evidence nodes, when true, provide support for the hypothesis. The next three are neutral, and the final three, when true, provide equally strong support against the hypothesis. As outlined above, this network is used to stochastically generate an evidence distribution for each model run. The hypothesis node is initialised as true with a probability determined by its base rate, and the evidence nodes are set to true with a probability that corresponds to their respective marginal conditional probabilities. Specifically, when using "big net", this means that H is true with a probability of 0.5, making it true in about half of the model runs. In the runs where H is true, evidence in favour of the hypothesis (i.e., more likely given H) is likely to be true, while evidence against it is more likely to be false (and vice versa). This generates an overall set of evidence values (of the form $E_1, \neg E_2, \ldots, \neg E_9$) that are in principle available to agents.

This results in a plausible evidence distribution in as much as there (likely) exists both evidence for and against the target hypothesis with the preponderance of that evidence varying across each run. Individual agents optimally combine whatever evidence they possess to calculate their current, posterior degree of belief in the hypothesis given their own evidence (and assuming the optimal prior, i.e., base rate with which the hypothesis will be true).

The 'best possible estimate' that agents could achieve under this set up, individually or collectively, corresponds to the posterior given all, in principle, available pieces of evidence in the simulated world, that is, the *optimal posterior* for a given run.

We used the model to simulate worlds in which agents are initialised with a random draw of either 1, 3, 5, or 7 pieces of evidence each, out of the possible total of 9. We simulated

200 instances of each for a total of 800 evidence distributions. We then identified, for each run, the lowest degree of belief amongst the agents in the set, the highest degree of belief given the initial evidence, the mean belief in the population of agents given their initial evidence, and the optimal posterior. Results are shown in Figure 3. The main take-away from these results is how likely it is that the optimal posterior is *more extreme* than the most extreme initial belief. This has the inevitable consequence that the collective mean is less extreme. As a consequence, averaging will inevitably pull a sizeable chunk of the population *away* from the best estimate.

Understanding better the reasons for this divergence provides further clues to how widespread this consequence might be, including in real world contexts. The discrepancy arises, fundamentally, because the integration of individual pieces of evidence gives rise to discontinuous jumps and changes in overall belief. Hearing a strong argument and revising one's beliefs accordingly may lead to large changes in belief. This means, for one, that the belief dynamics of simulated societies that exchange arguments are fundamentally different to those that will emerge in models based on averaging a single opinion (for extensive discussion, see Mäs & Flache, 2013; Assaad et al., 2023; Proietti & Chiarella, 2023). In particular, beliefs may readily become more extreme (either for the population as a whole, or for different subgroups) than the initial estimates—a dynamic that is antithetical to the notion of averaging itself (see Mäs & Flache, 2013; Hahn, 2023; Assaad et al., 2023).

To put it differently, the fundamental reason averaging fails is because the total evidential value of three independent pieces of evidence in favour of a claim is cumulative: each piece of evidence *adds* weight. The combined evidential value is not the average evidential value of the three individual arguments.

By the same token, the initial mean *would be* the optimal posterior (or very close to it) *only* if arguments for and against were roughly present in equal number and equal weight. For matters of fact, such a distribution will very much be the exception not the rule. This follows directly from a Bayesian perspective (see Hahn, 2023). However, it can readily be appreciated with an intuitive example as well: if it is, in fact, raining outside at present, it is extremely unlikely that there would be equal amounts of evidence speaking for and against the presence of rain. We consider evidence to be evidence precisely because it has some degree of lawful relationship to a hypothesis, so the true state of that hypothesis will, conversely, influence the (in principle) availability of evidence.

This means also that none of this hinges, fundamentally, on whether one is or is not adopting a Bayesian perspective, or modelling specifically Bayesian agents. It depends only on the way 'evidence' is related to the world and how multiple pieces of evidence plausibly combine. A pregnancy test, for example, is considered evidence for pregnancy precisely (and only) because it is more likely to indicate 'pregnant' than 'not pregnant' when used by a pregnant woman; and further positive tests increase our belief in pregnancy above and beyond just one. A Bayesian perspective is simply a particular formalisation of these more fundamental points.

Implications for the Utility of Averaging

All of this has implications for the utility of averaging as a means of collective decision-making, not just as a putatively rational strategy, but also as a practical tool for harnassing crowd wisdom for prediction and estimation.

If we consider again the sample run of Figure 2, averaging does improve accuracy. There is a 'wisdom of crowds effect' in as much as the squared error (i.e., the squared deviation from the optimal posterior) of the initial population mean is lower than the average accuracy across individuals in that population. By the same token, the average individual error decreases as the population converges. This is because squared error, like other proper scoring rules such as the logarithmic scoring rule (Carvalho, 2016), penalises the large distances of the beliefs on the far side of the population mean more than it rewards the short distances of the beliefs that are on the near side to the optimal value. The accuracy gains that result, however, seem limited in light of the overall error.

The discrepancy between the population average and the optimal posterior will, of course, decrease the more evidence individuals in the population have initially. Unfortunately, however, those will also be the circumstances in which the magnitude of the boost over average individual accuracy decreases.

To put all of this more generally, averaging boosts accuracy by eliminating variance. How useful that is depends on the source of that variance. Where it reflects random error, eliminating that variance will be extremely powerful and lead to estimates that are almost spookily accurate (Galton, 1907). Where that variance reflects unique knowledge, by contrast, the effects will likely be far less profound.

Crowd Wisdom and Communication

This, finally, has implications for research on crowd wisdom and the benefits (or harms) of communication. Motivated by formal results such as those discussed above (e.g., Equations 1-2), research has examined potential negative impacts of communication on crowd wisdom (Lorenz et al., 2011; Jönsson et al., 2015; Hahn, von Sydow, & Merdes, 2019; Zollman, 2010). What the discussion in this paper indicates clearly, however, is that it would be a mistake to use such results to caution against communication in contexts where crowd wisdom is the target more generally.

Specifically, the preceding discussions suggest that the nature of the estimate, and with it the nature of the variance, matters fundamentally. Where fragmentation of evidence is the underlying source of variance, promoting the exchange of reasons seems quite likely to outweigh potential costs of dependence. Experimental investigations with real communicating groups should seek to probe these boundary conditions further.

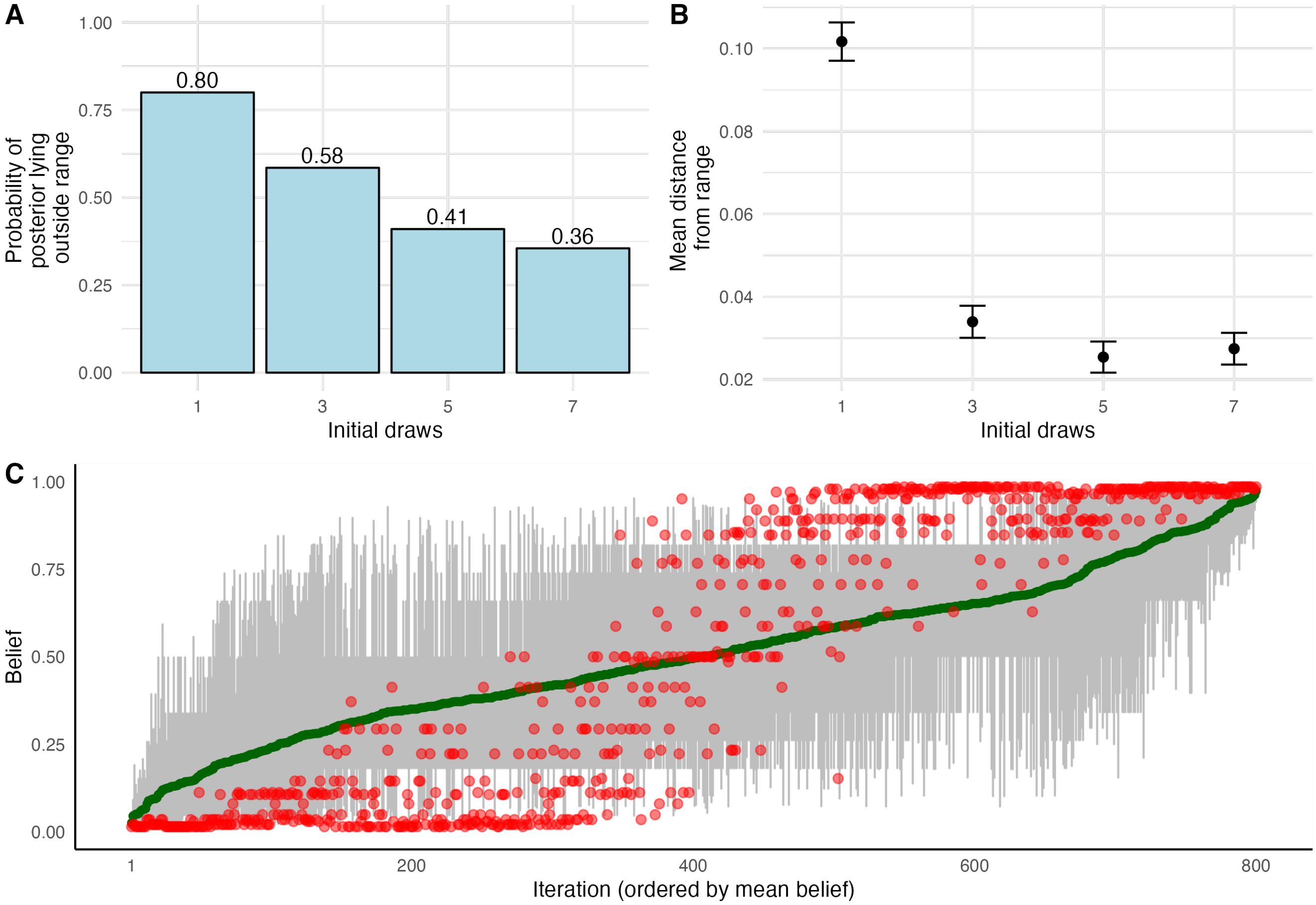


Figure 3: Results of simulations comparing optimal posteriors following argument exchange to agents' initial beliefs. **A.** The probability of the optimal posterior lying outside of the range of initial beliefs. Results are split by the number of initial draws parameter, which indicates whether the agents start with 1, 3, 5, or 7 pieces of initial evidence (out of a possible 9). **B.** Mean distance of the optimal posterior from the range of beliefs following averaging (i.e., how far below the minimum belief or how far above the maximum belief the optimal posterior is if it is outside the range of initial beliefs). Results are again split by the number of initial draws parameter. **C.** All 800 runs (iterations) of our model, ordered by the initial mean belief, with each iteration represented as a grey vertical bar (the range of initial beliefs), a green point (the mean initial belief), and red point (optimal posterior following argument exchange).

Conclusions

This paper compared argument exchange and opinion averaging as strategies of belief aggregation. We extended and analyzed an agent-based model (NormAN) of argument exchange and illustrated belief dynamics in comparison to a group of averaging agents. We then used simulations of initial belief distributions in that framework to illustrate the limits of averaging for attaining 'wisdom of crowd effects' and, by the same token, the limits of averaging models, given that (unrestricted) averaging will lead agents to converge toward the initial population mean. In circumstances where variance depends on differences in underlying information direct argument and evidence exchange facilitates convergence towards the optimal posterior belief. Opinion averaging, by contrast, will lead to an estimate that is insufficiently extreme. While averaging may enhance accuracy in groups with noisy, independent estimates, it fares poorly when agents' pre-deliberation beliefs are based on different pieces of specific evidence—all the more so if the body of available evidence unambiguously supports (or refutes) the hypothesis. These findings highlight the potential benefits of argument exchange on group accuracy and motivate further empirical and theoretical research into the merits and demerits of communication.

Acknowledgments

U.H. was supported by Arts and Humanities Research Council grant AH/V003380/1, and a Mercator Fellowship through Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 455912038. L.A. was supported by a Konrad-Adenauer-Stiftung scholarship.

References

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

Assaad, L., Fuchs, R., Jalalimanesh, A., Phillips, K., Schoeppl, L., & Hahn, U. (2023). A bayesian agent-based framework for argument exchange across networks. *arXiv preprint arXiv:2311.09254*.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Burton, J. W., Almaatouq, A., Rahimian, M. A., & Hahn, U. (2024). Algorithmically mediating communication to enhance collective decision-making in online social networks. *Collective Intelligence*, *3*(2), 26339137241241307.

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis*, *13*(4), 223 - 242.

Centola, D. (2018). *How behavior spreads: The science of complex contagions* (Vol. 3). Princeton University Press Princeton, NJ.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical association*, *69*(345), 118–121.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, *75*(7), 450–451.

Ghiselli, E. E. (1964). Theory of psychological measurement. *(No Title)*.

Hahn, U. (2022). Collectives and epistemic rationality. *Topics in Cognitive Science*, *14*(3), 602–620.

Hahn, U. (2023). Individuals, collectives, and individuals in collectives: the in-eliminable role of dependence. *Perspectives on Psychological Science*.

Hahn, U., Hansen, J. U., & Olsson, E. J. (2018). Truth tracking performance of social networks: how connectivity and clustering can make groups less competent. *Synthese*, 1–31.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in Cognitive Science*, *11*(1), 194–206.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, *5*.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*(1), 40–46.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, *114*(3), 733.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191–204.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856.

Ladha, K. K. (1992). The condorcet jury theorem, free speech and correlated votes. *American Journal of Political Science*, *36*(3), 617–634.

Lehrer, K., & Wagner, C. (1981). *Rational Consensus in Science and Society. A Philosophical and Mathematical Study*. Dordrecht: D. Reidel Publ. Co.

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

Mäs, M., & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PloS one*, *8*(11), e74516.

Page, S. (2006). The difference. *How the Power of Diversity Creates Better Groups, Firms*.

Proietti, C., & Chiarella, D. (2023). The role of argument strength and informational biases in polarization and bi-polarization effects. *Journal of Artificial Societies and Social Simulation*, *26*(2).

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of experimental Psychology*, *15*(5), 550.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business*. Doubleday New York.

Wallsten, T. S., Budescu, D. V., & Tsao, C. J. (1997). Combining linguistic probabilities. *Psychologische Beitrage*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks. *Nature*, *393*(6684), 440.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, *72*(1), 17–35.