Diversity and expertise in binary classification problems

Hein Duijf

Utrecht University

duijf.hein@gmail.com

Abstract

Democratic theorists and social epistemologists often celebrate the epistemic benefits of diversity. One of the cornerstones is the 'diversity trumps ability' result by Hong and Page (2004). Ironically, the interplay between diversity and ability is rarely studied in radically different frameworks. Although diversity has been studied in prediction and search problems, the diversity-expertise tradeoff has not been studied systematically for small, deliberative groups facing binary classification problems. To fill this gap, I will introduce a new evidential sources framework and study whether, when, and (if so) why diversity trumps expertise in binary classification problems.

1 Introduction

In an influential paper, Lu Hong and Scott Page (2004) introduce so-called landscape models and demonstrate that a functionally diverse group of agents can perform systematically better than a group of agents selected for their individual ability.[1] The idea that diverse teams are better problem-solvers than teams of experts is intuitively attractive. It feeds into an anti-epistocratic, egalitarian and deliberative-democratic narrative (Landemore, 2013): collective success does not come from hand-picking the best but from bringing a variety of skills to the table. It also suggests that science should promote diversity to optimize the pursuit of knowledge.

The diversity-expertise tradeoff concerns a possible tension between the individual expertise in a group and its diversity. Virtually everyone agrees that expertise requires epistemic competence (Hardwig, 1985; Goldman, 2001, 2018).[2] Accordingly, this paper focuses on this epistemic dimension and characterizes an agent's level of expertise as her capacity to adequately solve problems and reliably form true beliefs. Concepts of diversity can attach to different attributes such as gender, ethnicity, cognition, expertise, perspectives, heuristics, etc. Whether and how these different types of diversity can be valuable is an important research question (Steel et al., 2021; Sulik et al., 2022); nonetheless, following the large majority of work on the diversity-expertise tradeoff, this paper focuses on the value of cognitive diversity.

In Hong and Page's landscape models teams face a search problem where they must find the optimal solution among thousands of options. But what about binary classification problems where there are only two options available? Binary classification problems are common in legal, scientific, policy, and political domains. Consider a legal case where a group of judges has to decide on whether a defendant is guilty or not; a scientific case where a group of scientists has to figure out whether a particular virus causes some disease; a policy case where a committee has to decide on whether to approve or deny a given drug; or a political case where a group of political representatives must decide on whether all residents should be required to have health insurance. Landscape models cannot, and were never intended to, address these binary classification problems.

This limitation is not unique to landscape models, but characteristic of the broader literature on

[1] Also see the book-length discussion of related themes in (Page, 2007).

[2] In addition, virtually everyone agrees that moral character and social responsibility are characteristic of expertise and epistemic authority (Anderson, 2011; Rolin, 2021).

the wisdom of the crowds. Indeed, the literature predominantly focuses on aggregation (possibly preceded by deliberation) and on large groups, but the diversity-expertise tradeoff in small, deliberative groups has not been studied systematically for binary classification problems. This paper aims to fill this gap by introducing and utilizing a new modelling framework where agents deliberate to solve binary classification problems, where one option is objectively or intersubjectively correct.

I model deliberation as an exchange of 'evidences', such as opinions, empirical facts, observations, arguments, intuitions or personal experiences.[3] Evidences can be complex, such as personal experiences or complex arguments. Evidences might be (inferentially) misleading, that is, they can support the wrong option. The closest precursor is Dietrich and Spiekermann's (2025) model of deliberation as sharing and absorbing evidence, but they focus on the impact of pre-vote deliberation on the outcomes of majority voting. Another close relative is Hong and Page's (2025) recent model of interpreted signals where individuals use approximations (i.e., interpreted signals) to solve binary classification problems, but they focus on characterizing the range of collective accuracies under majority rule. Instead, my study aims to improve our understanding of the diversity-expertise tradeoff in *deliberation* — modelled as the exchange of (possibly complex, possibly misleading) evidences.

The paper is organized as follows. I will begin with presenting existing approaches to the phenomenon of the wisdom of the crowds, and show that there is a gap concerning the diversity-expertise tradeoff in binary classification problems (Sect. 2). Then, I will present the new evidential sources model (Sect. 3) and my simulation studies (Sect. 4). To illustrate the applicability of my analysis, I will briefly explore the ramifications for epistemic deliberative democracy (Sect. 5). I end with a brief summary and discussion (Sect. 6).

2 Diversity and ability in deliberative contexts

A group of individuals can outperform individual experts (de Condorcet, 1785; Galton, 1907), a phenomenon widely known as the wisdom of the crowds (Surowiecki, 2005; Page, 2007). To gain

[3] I follow the terminology of 'evidences' by Dietrich and Spiekermann (2025, 5): "Our notion of evidence is very broad and includes empirical facts as well as arguments, normative aspects, and other inputs into opinion formation."

a better understanding of collective wisdom and ignorance, it is important to draw a distinction between aggregation and deliberation, and to distinguish between different types of problems:[4]

- Prediction problems where agents must predict (or estimate or forecast) a certain value or quantity (e.g., the value of a stock);
- Search problems where agents must find the optimal solution among thousands of options (e.g., the best policy to a social problem);
- Sparse classification problems where agents must choose between a limited set of options (e.g., the best political candidate).

Prediction problems include Galton's (1907) famous problem of estimating the weight of an ox, where the average of the estimates of many individuals was more accurate than the estimates of individual experts. In these problems, the task is to determine a certain value or quantity such as the weight of an ox, the value of a stock, or the effective drug dose; which can be mapped to a continuous scale. Combining multiple predictions has been well studied (Armstrong, 2001). Moreover, not only empirical evidence (Batchelor and Dua, 1995) but also formal results (Lamberson and Page, 2012) indicate that diverse groups can trump expert teams. It should be noted, however, that the research on combining predictions concerns an aggregation problem and does not address the value of diversity in deliberation.

Search problems concern cases where one must find the optimal solution among many discrete options. In Hong and Page's (2004) seminal landscape models, teams deliberate to solve a search problem where they must find the optimal solution among thousands of options. One of their central results is a computational experiment where they report that random teams trumped teams composed of best-performing agents.

There have been several responses and modifications to the original work by Hong and Page.[5] Although their result has sometimes been taken to show that diversity always trumps ability, the actual results only demonstrate that diversity *can* trump ability in some circumstances. Indeed, partly because they only prove a possible link and not a necessary one, most of the ensuing literature has

[4] My typology is slightly more refined than the one by Landemore and Page (2015) of prediction and problem-solving tasks in that I refine their problem-solving tasks into search problems and sparse classification problems.

[5] Sakai (2020) provides an informative systematic literature review (up to 2020).

focused on explicating *when* diversity trumps ability.[6] In addition, Holman et al. (2018) and Grim et al. (2019) convincingly showed that Hong and Page's model does not adequately model expertise. They argue that an adequate model of expertise should include repeated success and show that experts in Hong and Page's model are not repeatedly successful. The exact details of their study need not concern us here; the only relevant point is that expertise should be modelled as repeated success.[7]

In any case, landscape models are limited to search problems and are ill-suited for studying binary or sparse classification problems. Indeed, Weymark (2015) proved that the landscape models are not applicable to binary problems. Moreover, I performed a simulation study that shows that landscape models are also ineffective for studying the tradeoff between diversity and expertise in sparse classification problems involving ten options.[8] These results are not surprising, since the landscape models were not meant to capture sparse classification problems. Rather, they indicate that we lack a framework that can inform us about the diversity-expertise tradeoff in deliberative contexts concerning sparse classification problems.

Sparse classification problems are cases where one must find the best option among a few options. In particular, in binary classification problems, agents must select the objectively or intersubjectively correct option among only two options. Condorcet famously showed that in such binary classification problems, under certain conditions, the majority judgment is more reliable than individual expert judgments (de Condorcet, 1785). Although some of these conditions have been questioned, there is, by now, a rich literature on jury theorems generalizing Condorcet's findings beyond these restricted settings (Dietrich and Spiekermann, 2023).

More recently, Dietrich and Spiekermann (2025) have utilized a new evidential sources' frame-

[6] For example, in the simulation studies by Holman et al. (2018) and Grim et al. (2019) diversity trumped expertise when the pool of conceptual resources is sufficiently wide. Reijula and Kuorikoski (2021) report that, in their study, diversity generally trumped expertise when the problem was complex, in the sense that it requires multiple component solutions and an efficient division of cognitive labour.

[7] I set aside the critique by Thompson (2014) that the results concern randomness (not diversity), since her criticism has been adequately addressed in (Kuehn, 2017; Singer, 2019). Furthermore, I want to acknowledge the recent critique by Genta (2024) that the result fails to be robust in ways that undermine its utility for justifying of deliberative democracy. However, for reasons of scope, extensive robustness tests of my evidential sources model are left for future work.

[8] I adopted the implementation by Huang (2024) and performed simulations for landscape models with 10 options. The unequivocal result is that diverse teams performed equally well as expert teams. My main reason for adopting Huang's implementation is that it was easily findable, accessible, and interoperable. I considered teams of size 9 in landscape models where $n = 10$, $k = 3$ and $l = 9$. The results were the same for smoothness factors ranging from 1 to 6 and trust factors 0, 0.33, 0.5, and 1 (see (Huang, 2024) for more details).

work to model deliberation and studied whether deliberation can improve majority decisions. Relatedly, Hong and Page (2025) have developed an interpreted-signals framework to characterize the range of collective accuracies under majority rule, which depends, among other things, on the diversity of the group. Going beyond the binary case, Keuschnigg and Ganser (2017) adopted the lens model and performed simulation studies to show that the effects of group composition depend on the social aggregation function and the type of problem: diversity is key mainly in prediction problems with aggregation via averaging and much less important in sparse classification problems with aggregation via plurality rule. In any case, the literature predominantly focuses on aggregation (possibly preceded by deliberation) and on large groups, but the diversity-expertise tradeoff in small, deliberative groups has not been studied systematically. This paper aims to fill this gap concerning the diversity-expertise tradeoff in deliberative contexts concerning binary classification problems.

3 The evidential sources model[9]

In this section, I will present the new evidential sources framework (Sect. 3.1), operationalize diversity within this framework (Sect. 3.2), and discuss some numerical examples and present some preliminary findings (Sect. 3.3).

Although the goal of this section is to describe the model's main components and mechanisms, a given model's explanatory value hinges on how well it matches the intended target. Although empirical correspondence is well beyond the scope of the paper, I would like to flag that one indirect way to gauge the empirical fit is to describe the epistemic competences of the pool of agents represented by a particular model (or parameter setting) and compare those to the real-world target. For example, given the political ignorance of voters (Somin, 2013), it seems reasonable that the pool of voters is best represented by models that yield pools of agents who are not very competent. From this perspective, let me foreshadow that I will describe the pools of agents that are represented by different parameter settings in my simulation study in Sect. 4.1 and will explore the implications for epistemic deliberative democracy in Sect. 5.

[9] All scripts necessary to reproduce the results are available as an anonymized GitHub repository including a walkthrough of the model on a GitHub page.

3.1 The model

My new modelling framework focuses on binary classification problems, where one option is objectively or intersubjectively correct. In these *evidential sources models*, agents can decide between two options by receiving information from sources. These sources can have different levels of reliability. For example, presumably, a reputable newspaper is a better source on political issues than a ten-year-old child; some scientific measurement tools are more accurate than others; and a court justice is more informed about the legal system than a 1st year philosophy student.

Models consist of sources, agents and teams. Figure 1 presents a stylized example with three agents and three sources, where edges represent the sources that are accessed by the agents. Let us discuss these in turn. *Sources* can be more or less reliable and some could even be systematically unreliable. The reliability of the sources is represented by probabilities. Each source s is assigned some probability p_s representing the likelihood that the source will generate, produce or share inferentially accurate evidence, i.e., evidence that supports the correct option. The 'valence' of an evidence measures which option it supports. We assume that the sources are probabilistically independent.

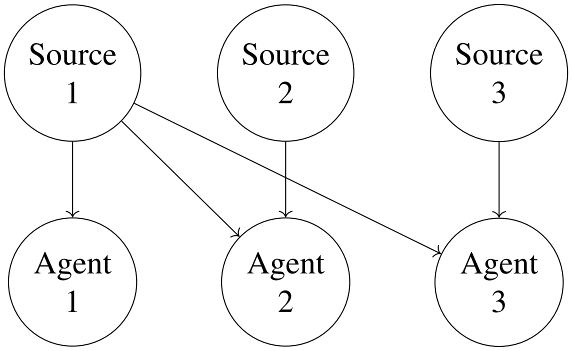

Figure 1: A stylized example.

Sources with reliabilities below 100% are imperfect in the specific sense that they may occasionally generate (inferentially) misleading evidence, i.e., evidence that supports the incorrect option. This imperfection is realistic, and it is compatible with malignant sources but also with truthful and sincere sources. After all, scientific evidence and inductive reasoning involves inductive risks. Hence, even truthful and sincere observations can be misleading, that is, support the incorrect option. Moreover, in non-monotonic reasoning contexts, facts can be misleading. Consider the classic example of Tweety. Suppose we were to learn the fact that Tweety is a bird. We would (legitimately) draw the conclusion that Tweety can fly. However, it turns out that Tweety is a penguin and is unable to fly. This case illustrates that truthful information (Tweety is a bird) can be misleading (i.e., support the conclusion that Tweety can fly).

An *agent's classification heuristic* is modelled by the sources she has access to. An agent forms her beliefs on the basis of the 'valence' of the evidences produced by the sources she has access to. For instance, in Figure 1, agent 2's heuristic is given by sources 1 and 2. Although sources were assumed to be independent, the fact that two agents' heuristics can overlap entails that agents need not be independent. Indeed, in Figure 1 the agents are not independent because they all access source 1.

For simplicity's sake, I assume that agents form their individual opinions by adopting the majority rule. This mechanism employs the simplifying assumption that all evidences have equal strength. This simplification offers an instructive, intelligible and tractable base case, which could be modified in future work.[10]

An agent's *individual score* represents their level of *expertise*; it is the probability that the agent succeeds in selecting the correct option. An agent's score can be calculated from the reliability of her sources. Indeed, we can compute the probability that the majority of evidences have the correct valence by going over all possible configurations of valences of the evidences of her sources. For example, in the simple case where all sources are equally reliable, the probability that k out of n evidences have the correct valence is binomially distributed, and we can easily compute the likelihood that the majority of the evidences have the correct valence.

My model of expertise is not susceptible to the criticism that affected Hong and Page's model where expertise failed to be correlated with repeated success (Holman et al., 2018; Grim et al., 2019). Let me explain. In my model, the individual score is the probability that an agent succeeds in selecting the correct option, averaging over all possible configurations of valences of the evidences of the sources (weighted by the probability of the configuration). In other words, experts in my model have the capacity to repeatedly utilize their sources to gather reliable evidences and form true beliefs.[11] Hence, higher individual scores represent higher likelihoods of repeated success.

[10] One possible extension could be that agents have (subjective or imperfect) assessments of the sources' relative reliability (see (Lehrer and Wagner, 1981; Klein and Sprenger, 2015)).

[11] Goldman (2018, 5) concurs that an expert with respect to a given domain "possesses more and/or better evidence

Although the model can allow for agents with heuristics of different sizes (i.e., one agent accessing ten sources and another agent accessing only one source), my simulation study focuses on agents with a fixed heuristic size.[12] The set of all possible heuristics (or agents) in a given model is determined by the number of available sources and the size of the heuristics. For example, when the model contains 21 available sources and the heuristic size is 5, there would be $\binom{21}{5} = 20,349$ possible heuristics.

Teams are modelled as sets of agents. Teams communicate internally before forming a collective decision between the two options. A team's *score* is the probability that it succeeds in selecting the correct option. I consider three deliberative mechanisms:

1. Opinion-based dynamics: agents communicate their opinion (not the underlying evidences) and the team's collective decision follows the majority rule;
2. Evidence-based dynamics: agents communicate all their evidences and the team's collective decision follows the majority rule; and
3. Boundedly rational evidence-based dynamics: agents communicate all their evidences and the team's collective decision follows the majority rule — possibly with double counting.

I will discuss these different deliberative mechanisms in more detail in Sect. 4. At the risk of oversimplification, we could say that the basic units in these mechanisms are opinions, evidences, and assertions, respectively. For example: in the first case, team members communicate their opinions concerning whether a given defendant is guilty of a robbery; in the second case, team members communicate their evidences, such as, an eyewitness statement or an argument that the defendant has no reason to *steal* the goods because they are wealthy; in the third case, team members not only communicate their evidences but an evidence also receives extra weight when it is communicated by multiple agents (in a sense, each separate assertion receives weight).

pertaining to propositions in [that domain] than most people in the relevant comparison class".

[12] This is broadly in line with the literature on landscape models, where simulation studies involve a fixed number steps and maximum step size in the agents' heuristics (resp., captured by parameters k and l). One notable exception is Genta (2024), who argues that Hong and Page's results fail to be robust when we consider pools of agents with a range of heuristic sizes.

3.2 Operationalizing diversity

Adapting from Hong and Page, the *diversity* of two heuristics A and B of the same size, notation: $\Delta(A, B)$, is defined by the proportion of sources that are in only one of them:

$$\Delta(A, B) = \frac{|A \setminus B|}{|A|}.$$

In other words, lower overlap between heuristics corresponds to higher diversity.[13]

A team $\mathcal{T}$ of heuristics is represented by a set of heuristics, i.e., a set of sets of sources ($\mathcal{T} \subseteq 2^S$). The diversity of a team $\mathcal{T}$ of at least two heuristics, notation: $\Delta(\mathcal{T})$, is given by the average diversity of pairs of heuristics in the team:

$$\Delta(\mathcal{T}) = \gamma \cdot \sum_{A,B \in \mathcal{T}} \Delta(A, B),^{14}$$

where $\gamma = \frac{1}{|\mathcal{T}|^2 - |\mathcal{T}|}$ is a normalization parameter so that $0 \leq \Delta(\mathcal{T}) \leq 1$.

Given an evidential sources model, a heuristic size, and a team size, one could enumerate all the possible teams and order them based on their diversity. Diverse teams are those teams that have the highest possible diversity.

3.3 Examples and preliminary findings

In this section, we discuss some numerical examples and draw out some preliminary findings and observations. First, in order to investigate the diversity-expertise tradeoff, we must consider situations where the heuristics are not equally good. Hence, we must consider evidential sources models where the reliabilities of the sources vary. After all, if all sources were equally reliable, all heuristics would be equally good.

Second, let us consider a numerical example where there are 18 available sources and the

[13] It may be helpful to note that $0 \leq \Delta(A, B) \leq 1$, and $\Delta(A, B) = 1$ iff $A \cap B = \emptyset$. When A and B have the same size, note that $\Delta(A, B) = \Delta(B, A)$, and $\Delta(A, B) = 0$ iff $A = B$.

[14] Singer (2019, 185) criticizes Hong and Page's notion of diversity and proposes a notion of "coverage diversity [which] measures how much of the heuristic space is covered by the group's combined heuristics." If transferred to the evidential sources model, coverage diversity of a team $\mathcal{T}$, $\Delta_C(\mathcal{T})$, would be defined by $\frac{|\bigcup \mathcal{T}|}{|S|}$, where S is the set of sources. However, this notion of coverage diversity is of limited use in the evidential sources framework because, in my simulation studies, the diverse team always has full coverage and the expert team has minimal coverage.

heuristic size is 3. Suppose that four sources are highly reliable (say, 90%) and all the others are unreliable (say, at chance level 50%). Let us consider teams of size 3. A diverse team could consist of agents who access only unreliable sources. The expert team, on the other hand, would consist of agents who only access highly reliable sources. Clearly, the expert team will perform much better than this diverse team — irrespective of the deliberative mechanism.

Third, observe that in the previous example, a diverse team could also consist of agents who access some highly reliable sources. More generally, any permutation of the sources will give rise to another team that is equally diverse. Hence, although there is typically only one expert team, there are multiple (maximally) diverse teams. The fact that there are multiple diverse teams complicates the analysis of the diversity-expertise tradeoff, because we must compare the performance of *the* expert team to the performance of a *set* of diverse teams. In rare cases, the expert team might trump some diverse teams but not others. To the best of my knowledge, this complication is rarely discussed in the literature on landscape models ((Singer, 2019) is a notable exception).

Fourth, let us consider another numerical example where there are 18 available sources and the heuristic size is 3. Suppose the sources are roughly equally reliable. In particular, suppose that four sources are slightly more reliable (say, 60%) than the others (say, 59%). Let us consider teams of size 3. In this case, the expert team would consist of agents who only access the slightly more reliable sources, and these experts will have a large degree of overlap. Diverse teams, on the other hand, would consist of agents without any overlap. As a result, it can easily be verified that diverse teams perform much better than the expert team — irrespective of the deliberative mechanism. The advantage of individual expertise is counteracted by the disadvantage of interdependence.

In sum, we observed that expertise trumps diversity in some cases, while diversity trumps expertise in others. Moreover, in virtually all cases, there are multiple (maximally) diverse teams, which complicates the analysis of the diversity-expertise tradeoff.

4 Simulation studies

In this section, I discuss and motivate the parameter choices in my simulation studies (Sect. 4.1) and then present the results of my simulation studies concerning opinion-based (Sect. 4.2), evidence-based (Sect. 4.3), and boundedly-rational evidence-based (Sect. 4.4) dynamics.

4.1 Simulation parameter choices

Simulation studies involve decisions that determine which areas of the model's parameter space are covered (and which are left out), because it is often impossible to cover the entire parameter space or because certain areas of the parameter space are implausible or irrelevant. I consider scenarios parametrized by the team size, the heuristics size, the number of available sources, and the sources' reliabilities. The goal of this subsection is to explain my parameter choices and to give an (indirect) interpretation of the different parameter settings.

In sum, my parameter decisions were:

1. Teams of size 9;
2. Heuristics of size 5;
3. Models with 13 and 17 available sources; and
4. Models with equidistant reliability distributions with means ranging from 0.55 to 0.75;

First, following the existing literature on landscape models, I decided to focus on teams of size 9 — although this decision is, admittedly, rather arbitrary.[15]

Second, people can only consider a limited number of evidences. Since empirical research in cognitive science strongly suggests that people can typically process and recall between 3 and 5 pieces of information (Cowan, 2001, 2010), I decided to only consider heuristics of size 5.[16][17] After all, the size of an agent's heuristic represents the number of evidences used in their individual belief-formation and decision-making.

[15] For example, Hong and Page (2004) report results for teams of size 10; and Grim et al. (2019) report results for teams of size 9 (and results for size 3 and 6 in their appendix).

[16] Odd heuristic sizes are helpful because they avoid ties.

[17] One might object that experts (or even laypeople) often have access to more than five sources. Let me briefly mention three responses: (1) In my view, the parameter choice is roughly compatible with an interpretation that agents have access to more than 5 sources but only 5 of those sources have a significant impact on their decision-making. (2) The adequate heuristic size ultimately depends on the intended real-world target, and the cited empirical evidence suggests that my modelling choice is likely appropriate in some circumstances. (3) I believe the adequacy of this parameter choice depends more on the epistemic competences of the pool of agents than on the (independent) plausibility that agents have access to a given number of sources. Notice that agents accessing only five sources can be highly competent: Figure 2 demonstrates that certain parameter settings in my simulation study correspond to situations where the pool of agents is quite expert. I thank Felix Kopecky and Max Noichl for pressing me on this point.

Third, the number of available sources is chosen in such a way as to correspond to the total number of available heuristics typically considered in simulation studies based on landscape models: between 1,320 and 6,840 available heuristics (Holman et al., 2018; Grim et al., 2019; Reijula and Kuorikoski, 2021).[18] Given my focus on heuristics of size 5, I considered models with 13 or 17 sources, which correspond to 1,287 and 6,188 possible heuristics, respectively.[19]

Fourth, my choice regarding the source reliability distributions is mainly guided by simplicity and tractability. I focused on, so-called, equidistant source reliability distributions that vary at most 0.2. These reliability distributions are represented by their mean. For example, mean 0.7 corresponds to an equidistant reliability distribution on the interval $[0.6, 0.8]$.[20]

I considered several such equidistant distributions with different means. I decided to include the equidistant distributions with means ranging from 0.55 to 0.75, in steps of 0.05. Since the best-performing heuristic has a score greater than 96% when the source reliability mean is 0.75, I did not consider higher means. Note that the equidistant distribution with mean 0.55 includes some anti-reliable sources whose reliability falls below chance level.

Given one such parameter setting, I considered all possible heuristics and computed their individual scores. The team of *experts* consisted of the heuristics with the highest individual scores. The *diverse* teams were generated by sequentially adding members using a greedy search procedure.[21] Roughly stated, at each step of this search procedure, a new heuristic that maximizes diversity is randomly selected. Consequently, I calculated the teams' scores by going over all possible configurations of valences of the evidences of the sources. The team's score is then given by the sum of the probabilities associated with those configurations that led to a correct team decision.[22]

[18] For landscape models, these correspond to $k = 3$ and $12 \leq l \leq 20$, respectively. As an exception, note that Hong and Page (2004) state that qualitatively similar results obtain for $2 \leq k \leq 7$, which would correspond to up to 390,700,800 heuristics.

[19] Note that the total number of possible heuristics is determined by the number of available sources and the heuristic size. So, if we take it seriously that the total number of possible heuristics should approximately be between 1,300 and 6,800, then that puts a constraint on the pairs of the number of available sources and the heuristic size. So, if the heuristic size should be 7 rather than 5, then that implies that the number of available sources is further restricted between (including) 13 and 15.

[20] For example, for 21 sources, this would correspond to the set of source reliabilities given by $\{0.6, 0.61, 0.62, \ldots, 0.79, 0.8\}$.

[21] Alternatively, one could consider all possible teams and compute their diversity. The diverse teams would then be the teams with the highest diversity. This is, however, computationally intractable because, for any of the considered evidential sources models there are more than 10^{22} possible teams.

[22] Notice that, when the team covers n sources, this analytical calculation must inspect 2^n configurations. So, for large sets of available sources, this may no longer be tractable.

Before proceeding to my simulation studies, I propose to improve our understanding of the parameter space in the simulation study. To do so, I present box plots of the individual scores of all possible heuristics in Figure 2, one for each parameter setting.[23] There are two useful ways to interpret this figure. On the one hand, these plots can inform us about the *difficulty* of the problems. After all, it seems plausible to say that lower individual scores are associated with harder problems. Indeed, in harder problems, everyone will perform worse. Hence, my simulation study might reveal whether diversity trumps expertise in hard or easy problems.

On the other hand, it seems plausible that these individual scores represent features of *the pool of agents*. Higher individual scores represent scenarios where the pool of agents has more knowledge and skills. The thought is that the individual scores do not only depend on the difficulty of the problem but also on the skill set of the particular sample of agents. For example, when deciding whether a defendant is guilty, the individual scores of legal experts will likely be higher than those of laypeople; and when determining whether a certain treatment will improve a patient's health, the individual scores of general practitioners will likely be higher than those of high-school students. The problem is the same, but the pool of agent differs. From this perspective, my simulation study might reveal whether diversity trumps expertise when the pool of agents is already quite expert or when it contains novices.

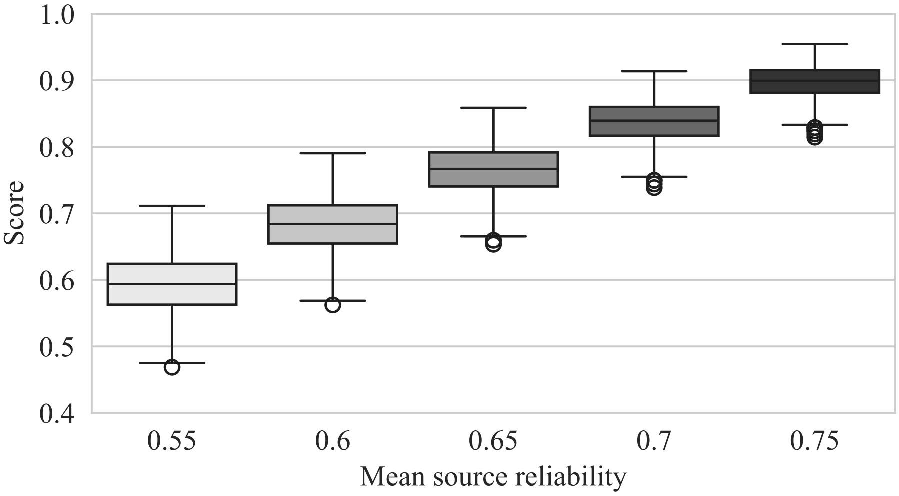

Figure 2: Individual scores.

[23] The figure shows the statistics for models with 13 sources; the statistics were virtually the same for models with 17 sources.

4.2 Simulation I: Opinion-based dynamics

In this section, I present the simulation results concerning opinion-based dynamics. In opinion-based dynamics, agents communicate their opinions — not the underlying evidences. Although this assumption may appear quite simplistic and unrealistic, many well-known models of belief propagation are opinion-based, where opinions are mapped to a continuous scale (DeGroot, 1974; Lehrer and Wagner, 1981; Hegselmann and Krause, 2002) or a nominal one (Liggett, 1985; Nowak et al., 1990; Axelrod, 1997). In any case, in my opinion-based dynamics, teams follow the majority opinion.

Figure 3 presents the results of my study, showing the performance difference between diverse teams and expert teams, that is, the difference between the average score of ten-thousand diverse teams and the expert team's score. It is instructive to consider both absolute and relative performance differences. To report relative differences, I decided to report the relative difference in the *error rates* of diverse and expert teams. This relative difference is measured relative to the team that performs best in the given circumstances. For example, if the expert team's score is 80% and the diverse team's score is 90%, their error rates would be 20% and 10%, respectively, the difference 10%, and the relative difference 100%.

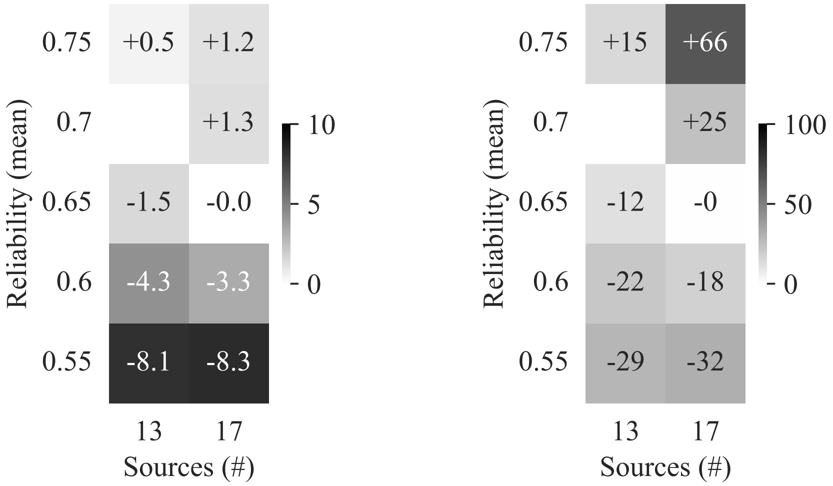

Figure 3: The diversity-expertise tradeoff for opinion-based dynamics. Positive numbers indicate parameter settings where diversity trumped expertise; and negative numbers represent those where expertise trumped diversity. Numbers represent the absolute performance difference (left) and relative difference (right) in percentages. Results are masked if (and only if) $p \geq .001$ (one-sample Wilcoxon test).

The results clearly indicate that there is no universal winner: diverse teams outperform expert teams in some contexts, and vice versa. The tradeoff appears to depend on the number of available sources and especially their reliability. Let me summarize some observations. First, diversity trumped expertise when sources where highly reliable. After all, diverse teams performed better than expert teams when the mean reliability was 70 or 75 percent.

Second, expertise is more valuable if sources can be unreliable (i.e., close to chance level) or anti-reliable (i.e., below chance level). Indeed, expertise trumped diversity when the mean source reliability was 60% and 55%, i.e., when sources could be unreliable or anti-reliable, respectively. Moreover, when sources are weakly reliable (i.e., 65%), expert teams continued to perform slightly better than diverse teams. Hence, my simulation study showed that expertise is valuable when sources can be unreliable or anti-reliable. Moreover, given the observations at the end of Sect. 4.1, diversity trumped expertise mainly when the problem was easy and/or the pool of agents was already quite expert.

Third, classification problems can afford broader or narrower pools of conceptual resources. If we think that the pool of conceptual resources is correlated with the number of available sources, then the simulation results appear to indicate that, when the sources are not anti-reliable, diversity becomes more valuable when the pool of conceptual resources is widened. This partially confirms Grim et al.'s finding that wider pools of conceptual resources benefit diversity; however, contrary to their findings, the width of the pool of conceptual resources did not affect the qualitative outcome of whether diversity trumped expertise in my study.

Before moving on to further simulation studies, I should remark that *statistical* significance is different from *practical* significance. So, although most simulation results were statistically significant, we ought to be careful in drawing conclusions about their practical significance. On the one hand, it often seems questionable whether an absolute performance difference of, say, 1% bears any practical significance. In some cases, this tiny performance advantage might be outweighed by other considerations. On the other hand, in high stakes decisions, a 1% (absolute) performance improvement might be highly relevant.

What explains our results? The diversity-expertise tradeoff concerns the tradeoff between selecting for ability or diversity. Let us start with ability. One would expect that higher individual performance differences favour the expert team. Figure 4 presents the average difference in individual scores between members of the expert team and the diverse team, showing both absolute and relative performance differences. Observe that the absolute individual differences decrease when the source reliabilities increase, while the relative individual differences increase. Hence, the absolute measure provides the best insight into the diversity-expertise tradeoff.

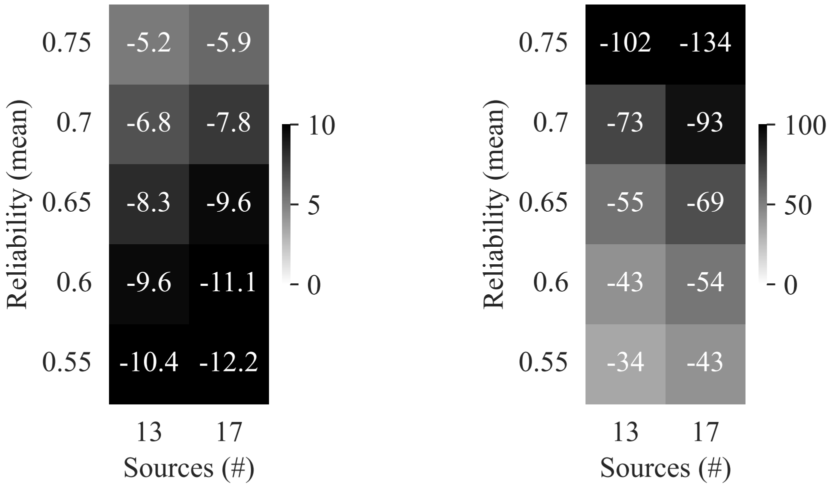

Figure 4: The performance difference of individual team members. Numbers represent the absolute performance difference (left) and relative difference (right) in percentages. All results were statistically significant (one-sample Wilcoxon test, $p < .001$).

We proceed with diversity. The difference in diversity between the expert and the diverse team was 0.41 and 0.51 for scenarios with 13 and 17 sources, respectively. For example, when there were 13 sources available, the difference in diversity implied that the members of the expert team on average had 41% *more overlap* than the members of the diverse team. In particular, given the heuristic size of 5, this meant that, on average, members of the expert team had 2 out of 5 sources more in common compared with members of the diverse team. To grasp the extent of the interdependence between experts, we can show that *disagreement between experts is highly unlikely*. Indeed, the likelihood that two experts disagree varies between 6% and 25% and decreases when sources become more reliable (see Table 2).

The high degree of interdependence between experts prompts the hypothesis that the expert team may not do much better than the highest-scoring individual. Indeed, we can show that the expert team performs only slightly better than the highest-scoring individual: the absolute performance improvement is at most 1.7% (see Table 2). Moreover, in all scenarios where the expert *team* trumped diverse teams, the highest-scoring *individual* also trumped diverse teams.

In conclusion, the difference in individual performance and low rate of expert disagreement jointly explain the diversity-expertise tradeoff for the opinion-based dynamics. The homogeneity of the expert team makes it the case that it does only slightly better than the highest-scoring individual. In scenarios where the individual performance difference was large enough (i.e., bigger than 8%), this lack of diversity among experts did not jeopardize their lead over teams of diverse — but worse-performing — individuals. By contrast, when the individual performance difference was smaller (i.e., below 8%), diversity trumped expertise.

4.3 Simulation II: Evidence-based dynamics

The difference between the opinion-based dynamics and the evidence-based ones is in the content of the communication. In opinion-based dynamics, agents communicate their opinions, while the agents communicate all their evidences in evidence-based dynamics. Similar evidence-based dynamics have been introduced before (e.g., Mäs and Flache, 2013; Ding and Pivato, 2021; Dietrich and Spiekermann, 2025) but none of them addressed the diversity-expertise tradeoff. Although we previously observed that expertise often trumped diversity for the opinion-based dynamics, it is plausible to think that the *evidence-based* dynamics favours diversity. After all, diverse teams have a greater coverage and will therefore share more evidences. When evidences are generally inferentially reliable, it seems plausible that greater diversity will lead to better team decision-making.

Let us consider the results of my simulation study concerning evidence-based dynamics: see Figure 5. Let me sum up a few observations. First, to my surprise, the *qualitative* results regarding the diversity-expertise tradeoff are virtually identical to the opinion-based dynamics. That is, I expected there to be scenarios with a strong *DTE reversal*: that is, scenarios where expertise trumped diversity for opinion-based dynamics while diversity trumped expertise for evidence-based dynamics. However, a comparison of Figures 3 and 5 reveals that, except for the scenario where there were 17 sources available with a mean reliability of 65%, such diversity-expertise reversals did not obtain in the considered scenarios. Hence, once again, diversity trumped expertise mainly when sources were (at least weakly) reliable, the problem was easy, and/or when the pool of agents was already quite expert.

Second, the quantitative results appear to indicate that the evidence-based dynamics favour

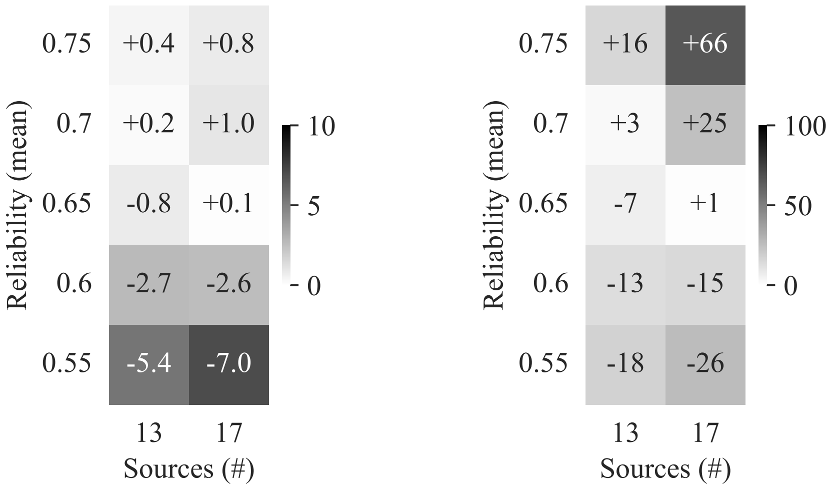

Figure 5: The diversity-expertise tradeoff for evidence-based dynamics. Positive numbers indicate parameter settings where diversity trumped expertise; and negative numbers represent those where expertise trumped diversity. Numbers represent the absolute performance difference (left) and relative difference (right) in percentages. All results are statistically significant (one-sample Wilcoxon test, $p < .001$).

diversity over expertise, in comparison with opinion-based dynamics. Indeed, the absolute difference shifted slightly towards diversity, at least in absolute terms (except when the mean source reliability was 75%). In contrast, the *relative* difference hardly changed between opinion-based and evidence-based dynamics. This suggests that evidence-based dynamics had an approximately similar relative effect on the performance of expert and diverse teams, while having a disparate absolute effect.

What explains this diversity-expertise tradeoff? In evidence-based dynamics, the diverse team acts like an individual who accesses all sources. In contrast, the expert team acts like an individual who accesses the top sources — but not all sources. More specifically, in the considered scenarios, the expert team acted like an individual who accesses the top 9 sources. So, the diverse team covered 4 and 8 sources more than the expert team when there were 13 and 17 sources available, respectively. Hence, the diversity-expertise trade-off is constituted by the advantage of including more sources and the disadvantage of including suboptimal sources. For example, when the source reliability mean was 55%, some sources were anti-reliable and, hence, it is not surprising that expertise trumped diversity. In comparison, it is somewhat surprising that the expert team performed better than the diverse team when there were 13 sources available and the source reliability mean was 65%. After all, in that scenario, all sources were (at least weakly) reliable (above 55%). In sum, the diversity-expertise tradeoff for evidence-based dynamics is mainly explained by the sources' reliabilities.

My analysis triggers some further hypotheses that could be studied in future work. Although all the considered models had the same source reliability variance, the framework can be utilized to investigate the hypothesis that increasing the *variance* will favour expertise. In addition, although the team size was fixed, larger expert teams will have higher coverage while diverse teams already had maximum coverage. It thus appears that diverse and expert teams will become more similar as the team size increases.

4.4 Simulation III: Boundedly rational evidence-based dynamics

The previous evidence-based dynamics could be called unboundedly rational in the sense that the team's decision procedure avoids overcounting evidences. In the boundedly-rational dynamics, when two agents communicate the same evidence, that evidence receives extra weight. Hence, evidences can be overcounted, which exaggerates the influence of that evidence and may lead to suboptimal collective decision-making. Overcounting aligns with the empirical finding that people believe repeated information more than new information, a phenomenon known as the repetition-induced truth-effect (Hasher et al., 1977; Dechêne et al., 2010). Moreover, overcounting is not necessarily irrational, especially when certain evidences are more reliable than others or when certain sources are more trustworthy or reliable than others (Lehrer and Wagner, 1981).[24]

It is hard to form a hypothesis for these dynamics. On the one hand, the expert team seems to suffer heavily from overcounting, given the higher degree of overlap. Diverse teams, given their wide coverage and minimal overlap, will not suffer as much from overcounting evidences: all sources will be counted roughly equally. Thus, prima facie, expert teams seem more vulnerable to *skewed* overcounting. On the other hand, it seems plausible that overcounting is only suboptimal

[24] Another potential way to interpret this deliberative mechanism is that the members communicate their opinion and their confidence, where their confidence is understood as the valences of their evidence. For example, when an agent received four evidences with a valence supporting the correct option and one supporting the incorrect opinion, they would communicate this information without explicating the exact nature and content of their evidences. Accordingly, teams would be unable to determine the overlap in the evidences and, hence, the boundedly rational evidence-based dynamics seems plausible and justifiable for this model of group communication. I thank Dominik Klein for this suggestion.

when it leads to overcounting sources *regardless of their reliability*. Given that the expert team is overcounting the most reliable sources, overcounting appears to track the fact that certain sources are more reliable. By contrast, since diverse teams are not necessarily utilizing the most reliable sources, overcounting appears to be divorced from the sources' reliability. Hence, diverse teams appear to be more vulnerable to *arbitrary* overcounting.

Let us see how these opposite predictions pan out by considering the results presented in Figure 6. Let me sum up a few observations. First, again to my surprise, the *qualitative* results regarding the diversity-expertise tradeoff are virtually always irrespective of the deliberative mechanism — the only exception is the scenario with 17 sources and a mean reliability of 65%.

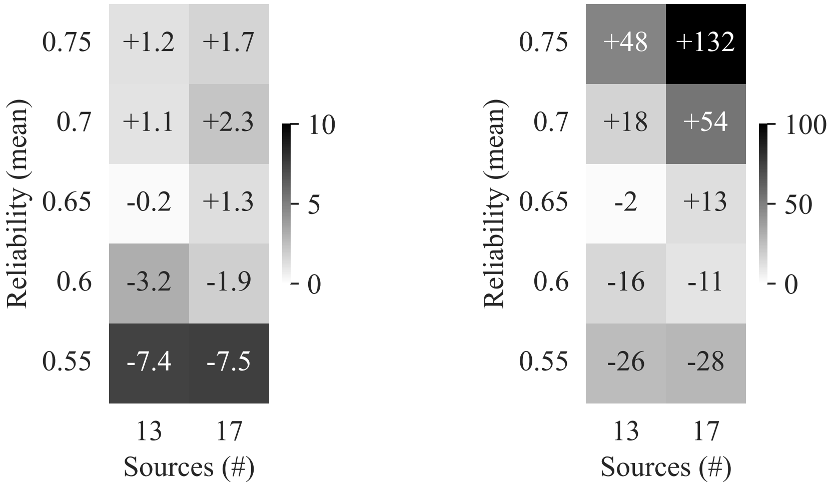

Figure 6: The diversity-expertise tradeoff for boundedly rational evidence-based dynamics. Positive numbers indicate parameter settings where diversity trumped expertise; and negative numbers represent those where expertise trumped diversity. Numbers represent the absolute performance difference (left) and relative difference (right) in percentages. All results are statistically significant (one-sample Wilcoxon test, $p < .001$).

Second, in scenarios where diversity trumped expertise, overcounting further expanded the performance advantage of diverse teams (compared with the other deliberative mechanisms). In contrast, in scenarios where some sources were anti-reliable, overcounting increased the comparative deficit of diverse teams. Hence, the risk of arbitrary overcounting in diverse teams appears to be outweighed by the risk of skewed overcounting in expert teams when sources are not unreliable.

What explains these results? The expert team heavily overcounted the top sources (see Table 1). Hence, it appears that the expert team may act similar to the best-performing individual. Indeed, we can show that the expert team performed only slightly better: the (absolute) improvement is at most 1.6% (see Table 2). Nonetheless, whenever the expert team trumped diverse teams, the best-performing *individual* also trumped diverse teams. So, the qualitative results concerning the diversity-expertise tradeoff appear to hinge on the best-performing individual.

Source number	1	2	3	4	5	6	7	8	9
Source weight	9	9	8	6	5	3	2	2	1

Table 1: Source weights for the expert teams.

In comparison, the diverse team did not heavily overcount any sources — all sources were counted twice or thrice. So, it seems plausible that diverse teams acted similarly for both evidence-based dynamics. Indeed, we can show that there is only a minute performance difference of at most 0.5% (see Table 2). This confirms the suspicion that overcounting hardly affected diverse teams.

All in all, for boundedly rational evidence-based dynamics, diverse teams acted similar to diverse teams with evidence-based dynamics, while expert teams acted similar to the best-performing individual. So, the diversity-expertise tradeoff is, once again, constituted by the advantage of including more sources and the disadvantage of including suboptimal sources. However, in contrast with the evidence-based dynamics where the expert team acted like an individual who accesses the top 9 sources, in this case, the expert team acted similar to an individual who accesses the top 5 sources. Hence, for the boundedly rational evidence-based dynamics, I suspect that the expert team's score will not significantly increase when the team is enlarged.

5 Epistemic deliberative democracy

Epistemic democrats consider the epistemic value of democracy (Cohen, 1986; Estlund, 2008; Estlund and Landemore, 2018; Goodin and Spiekermann, 2018; Landemore, 2013). The basic assumption of epistemic democrats is that there exists some procedure-independent measure of success (i.e., truth), and democratic decision procedures should be evaluated in terms of their reliability of leading to successful outcomes, by that measure (Cohen, 1986). Optimistic epistemic democrats often rely on mathematical results such as Condorcet's jury theorems, the miracle of aggregation or Hong and Page's diversity-trumps-ability result to argue that democratic decision procedures are epistemically valuable.

These epistemic defences of democracy are threatened by the empirical fact that citizens are far from perfect in assessing which of several policy proposals or political representatives is best. Political scientists have documented a staggering lack of political knowledge (Achen and Bartels, 2017; Somin, 2013). But, voter knowledge is not the same as voter competence and voters can and do successfully use informational cues and shortcuts in their political decision-making (Lupia, 2006; Goren, 2013). Moreover, comparisons between uninformed voters and their informed counterparts show that uninformed voters fall between a coin toss and the informed voter (Bartels, 1996; Lau and Redlawsk, 1997, 2006). Hence, under the weak assumption that informed voters are better than chance level, the empirical evidence indicates that voters are minimally competent in the sense that they are better than chance level at determining which political option is best. Although minimal competence may suffice to defend democratic *aggregation* methods (e.g., majority voting) on the basis of Condorcet-like jury theorems (Goodin and Spiekermann, 2018), it is an open question whether it suffices for *deliberative* democracy.

What do my simulation results entail for the epistemic value of deliberative democracy? I will assume that deliberative democracy is operationalized in terms of small, deliberative citizen assemblies, and I will adopt the assumption of epistemic democrats that there exists some procedure-independent measure of success. To apply my simulation results to this context, we must determine which parameter settings best capture these deliberative assemblies. Although rigorously fitting the model's parameters lies beyond the scope of the current paper, we can make an informed estimate. Since uninformed voters fall between a coin toss and the informed voter (Bartels, 1996; Lau and Redlawsk, 1997, 2006), it is plausible that the average voter competence is most likely below 80% and probably even below 70%. Hence, Figure 2 illustrates that we should focus on evidential sources models with mean source reliability of 55% and 60% (and, more optimistically, perhaps 65%). My simulation study showed that expert teams trump diverse teams in these circumstances. So, my results undermine epistemic arguments in favour of deliberative democracy. Moreover, this conclusion holds robustly for the three considered deliberative mechanisms. Hence, from a purely epistemic perspective, when selecting for small, deliberative assemblies, one should prioritize expertise over diversity.[25]

[25] Of course, these epistemic considerations may be overruled by other non-epistemic considerations, such as po-

6 Conclusion

The idea that promoting diversity leads to better collective decision-making is a powerful one. However, there is a tension between selecting for diversity or ability. On the one hand, diverse teams can make use of more diverse evidences, but, without appropriate epistemic guardrails, they are at a risk of relying on suboptimal (or even unreliable) evidences. On the other hand, expert teams leverage the most reliable evidences at the expense of lower evidential coverage, excluding many (potentially) available evidences. I have shown that diversity and expertise are important for optimizing collective decision-making in different circumstances, where diversity enhances epistemic coverage and expertise ensures deliberation is based on reliable information.

In my study, diversity trumped expertise in some situations, and vice versa. I presented a novel evidential sources framework and analysed whether, when and (if so) why diversity trumps expertise for three different deliberative mechanisms — one based on opinions and two on evidences. Surprisingly, the qualitative results were virtually identical for all three deliberative mechanisms. Generally, expertise trumped diversity when some sources were anti- or unreliable, when the problem was hard, and/or when the pool of agents contained novices. In contrast, diversity trumped expertise when sources were (at least weakly) reliable, when the problems were easy, and/or when the pool of agents was already quite expert.

What explained the diversity-expertise tradeoff? On the one hand, in opinion-based dynamics, the tradeoff could be explained by the difference in individual performance and the rate of expert disagreement. When the differences were small and/or the disagreement rate was low, diversity trumped expertise. On the other hand, in evidence-based dynamics, the tradeoff was constituted by the advantage of including more sources and the disadvantage of including suboptimal (or unreliable) sources. Given the better coverage of diverse teams, this only translated into an advantage if and when sources were (at least weakly) reliable. All in all, then, diversity and expertise are both important for addressing the risks of suboptimal coverage and unreliable evidence, respectively.

Overall, my results suggest that optimizing the team composition will require selecting for diversity with some epistemic guardrails. We must ensure that weak evidences are excluded and epistemic coverage is optimized. Hence, the key aim is to improve collective knowledge acquisi-

litical or moral considerations.

tion by striking the right balance between expertise and diversity.

A Appendix

Number of sources	Source reliability (mean)	(a)	(b)	(c)	(d)
13	0.55	24.7	0.7	0.7	0.10
	0.60	21.0	1.3	1.2	0.17
	0.65	16.5	1.5	1.5	0.18
	0.70	11.6	1.4	1.4	0.15
	0.75	7.1	1.1	1.0	0.09
17	0.55	24.0	1.1	1.1	0.31
	0.60	20.0	1.6	1.5	0.49
	0.65	15.4	1.7	1.6	0.46
	0.70	10.6	1.5	1.4	0.32
	0.75	6.2	1.0	1.0	0.16

Table 2: (a) The (maximum) rate of disagreement among experts; the performance difference between the expert team and the best-performing individual for opinion-based dynamics (b) and boundedly-rational evidence-based dynamics (c); and (d) the performance difference between diverse teams for both evidence-based dynamics (all the results were statistically significant (paired Wilcoxon test, $p < .001$)). All in percentage.

References

Achen, C. H. and L. M. Bartels (2017, August). *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton University Press.

Anderson, E. (2011, June). Democracy, public policy, and lay assessments of scientific testimony. *Episteme 8*(2), 144–164.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pp. 417–439. Boston, MA: Springer US.

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution 41*(2), 203–226.

Bartels, L. M. (1996). Uninformed votes: Information effects in presidential elections. *American Journal of Political Science 40*(1), 194–230.

Batchelor, R. and P. Dua (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science 41*(1), 68–75.

Cohen, J. (1986, October). An epistemic conception of democracy. *Ethics 97*(1), 26–38.

Cowan, N. (2001, February). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences 24*(1), 87–114.

Cowan, N. (2010, February). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science 19*(1), 51–57.

de Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.*

Dechêne, A., C. Stahl, J. Hansen, and M. Wänke (2010, May). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review 14*(2), 238–257.

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

Dietrich, F. and K. Spiekermann (2023). Jury Theorems. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University.

Dietrich, F. and K. Spiekermann (2025, March). Deliberation and the wisdom of crowds. *Economic Theory 79*(2), 603–655.

Ding, H. and M. Pivato (2021, May). Deliberation and epistemic democracy. *Journal of Economic Behavior & Organization 185*, 138–167.

Estlund, D. (2008). *Democratic Authority: A Philosophical Framework.* Princeton University Press.

Estlund, D. and H. Landemore (2018). *The Epistemic Value of Democratic Deliberation*, pp. 112–131. Oxford University Press.

Galton, F. (1907, March). Vox populi. *Nature 75*(1949), 450–451.

Genta, B. S. (2024, October). Formal models and justifications of democracy. *Synthese 204*(5), 138.

Goldman, A. I. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research 63*(1), 85–110.

Goldman, A. I. (2018, March). Expertise. *Topoi 37*(1), 3–10.

Goodin, R. E. and K. Spiekermann (2018). *An Epistemic Theory of Democracy*. Oxford: Oxford University Press.

Goren, P. (2013). *On Voter Competence*. Oxford: Oxford University Press.

Grim, P., D. J. Singer, A. Bramson, B. Holman, S. McGeehan, and W. J. Berger (2019, January). Diversity, ability, and expertise in epistemic communities. *Philosophy of Science 86*(1), 98–123.

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy 82*(7), 335–349.

Hasher, L., D. Goldstein, and T. Toppino (1977, February). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior 16*(1), 107–112.

Hegselmann, R. and U. Krause (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation 5*(3).

Holman, B., W. J. Berger, D. J. Singer, P. Grim, and A. Bramson (2018). Diversity and democracy: Agent-based modeling in political philosophy. *Historical Social Research / Historische Sozialforschung 43*(1), 259–284.

Hong, L. and S. E. Page (2004, November). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences 101*(46), 16385–16389.

Hong, L. and S. E. Page (2025, February). The range of collective accuracy for binary classifications under majority rule. *Economic Theory 79*(1), 275–300.

Huang, A. C. W. (2024, July). Landscapes and bandits: A unified model of functional and demographic diversity. *Philosophy of Science 91*(3), 579–594.

Keuschnigg, M. and C. Ganser (2017). Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Management Science 63*(3), 818–828.

Klein, D. and J. Sprenger (2015). Modelling individual expertise in group judgements. *Economics & Philosophy 31*(1), 3–25.

Kuehn, D. (2017, January). Diversity, ability, and democracy: A note on Thompson's challenge to Hong and Page. *Critical Review 29*(1), 72–87.

Lamberson, P. J. and S. E. Page (2012). Optimal forecasting groups. *Management Science 58*(4), 805–810.

Landemore, H. (2013). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton: Princeton University Press.

Landemore, H. and S. E. Page (2015, August). Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, Philosophy & Economics 14*(3), 229–254.

Lau, R. R. and D. P. Redlawsk (1997). Voting correctly. *The American Political Science Review 91*(3), 585–598.

Lau, R. R. and D. P. Redlawsk (2006). *How Voters Decide: Information Processing during Election Campaigns*. Cambridge Studies in Public Opinion and Political Psychology. Cambridge ; New York: Cambridge University Press.

Lehrer, K. and C. Wagner (1981). *Rational Consensus in Science and Society*. Dordrecht: D. Reidel Publishing Company.

Liggett, T. M. (1985). *Interacting Particle Systems*. Berlin/Heidelberg: Springer.

Lupia, A. (2006, January). How elitism undermines the study of voter competence. *Critical Review 18*(1-3), 217–232.

Mäs, M. and A. Flache (2013, November). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLOS ONE 8*(11), e74516.

Nowak, A., J. Szamrej, and B. Latané (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review 97*(3), 362–376.

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

Reijula, S. and J. Kuorikoski (2021, December). The diversity-ability trade-off in scientific problem solving. *Philosophy of Science 88*(5), 894–905.

Rolin, K. H. (2021, December). Objectivity, trust and social responsibility. *Synthese 199*(1), 513–533.

Sakai, R. (2020, June). Mathematical models and robustness analysis in epistemic democracy: A systematic review of diversity trumps ability theorem models. *Philosophy of the Social Sciences 50*(3), 195–214.

Singer, D. J. (2019, January). Diversity, not randomness, trumps ability. *Philosophy of Science 86*(1), 178–191.

Somin, I. (2013). *Democracy and Political Ignorance*. Stanford: Stanford University Press.

Steel, D., S. Fazelpour, B. Crewe, and K. Gillette (2021, February). Information elaboration and epistemic effects of diversity. *Synthese 198*(2), 1287–1307.

Sulik, J., B. Bahrami, and O. Deroy (2022, May). The diversity gap: When diversity matters for knowledge. *Perspectives on Psychological Science 17*(3), 752–767.

Surowiecki, J. (2005, August). *The Wisdom of Crowds*. Knopf Doubleday Publishing Group.

Thompson, A. (2014). Does diversity trump ability? *Notices of the AMS 61*(9), 1024–1030.

Weymark, J. A. (2015). Cognitive diversity, binary decisions, and epistemic democracy. *Episteme 12*(4), 497–511.