Philosophy of Science (2026), **93**, 631–654
doi:10.1017/psa.2026.10200

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Diversity and Expertise in Binary Classification Problems

Hein Duijf

Department of Philosophy and Religious Studies, Utrecht University, the Netherlands
Email: h.w.a.duijf@uu.nl

(Received 12 August 2025; revised 12 December 2025; accepted 10 February 2026; first published online 26 February 2026)

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. In particular, the diversity–expertise trade-off 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. The newly gained insights are used to revisit the epistemic credentials of deliberative democracy.

1. Introduction

The idea that diverse teams are better problem-solvers than teams of experts is intuitively attractive. 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] Their "diversity trumps ability" result is one of the cornerstones of 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.[2] Ironically, however, the diversity–expertise trade-off in deliberative collectives has rarely been studied in radically different frameworks. This paper aims to fill this important gap, and revisit the implications for epistemic deliberative democracy.

The diversity–expertise trade-off concerns a possible tension between the individual expertise in a group and its diversity. Virtually everyone agrees that

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

[2] I should flag that Hong and Page's "diversity trumps ability" result has been criticized and contextualized. Important contributions include those by Thompson (2014), Kuehn (2017), Singer (2019), Holman et al. (2018), Grim et al. (2019), Reijula and Kuorikoski (2021), and Genta (2024).

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

expertise requires epistemic competence (Hardwig 1985; Goldman 2001, 2018).[3] 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 trade-off, this paper focuses on the value of cognitive diversity, which refers to diverse perspectives, interpretations, heuristics, and predictive models (e.g., Landemore 2013, 89).[4]

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 deliberative groups facing binary classification problems.

This limitation is not unique to landscape models, but characteristic of the broader literature on the wisdom of the crowd. Indeed, the literature predominantly focuses on aggregation (possibly preceded by deliberation) and on large groups, but the diversity–expertise trade-off 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 modeling 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.[5] 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 trade-off in *deliberation*—modeled as the exchange of (possibly complex, possibly misleading) evidences.

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

[4] See section 3.2 for more details about the concept of diversity at the heart of this paper.

[5] 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."

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

In turn, I will argue that my approach provides new insights on the epistemic credentials of *deliberative* democracy.

The paper is organized as follows. I begin with presenting existing approaches to the phenomenon of the wisdom of the crowd, and show that there is a gap concerning the diversity–expertise trade-off for deliberative groups facing binary classification problems (section 2). Then, I present the new evidential sources model (section 3) and my simulation studies (section 4). Subsequently, I explore the ramifications for epistemic deliberative democracy (section 5). I end with a brief summary and discussion (section 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 crowd (Surowiecki 2005; Page 2007). Diversity is often celebrated as an important driver of the wisdom of the crowd. To gain 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:[6]

- 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, that 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 the best-performing agents.

[6] 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.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

There have been several responses and modifications to the original work by Hong and Page.[7] 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 focused on explicating *when* diversity trumps ability.[8]

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 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 trade-off between diversity and expertise in sparse classification problems involving ten options.[9] These results are not surprising, since the landscape models were not meant to capture sparse classification problems. Rather, they indicate that landscape models cannot inform us about the diversity–expertise trade-off 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 framework 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 trade-off in small, deliberative groups has not been studied systematically. This paper aims to fill this gap.

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

[8] 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 labor.

[9] I adopted the implementation by Huang (2024) and performed simulations for landscape models with ten options. The unequivocal result is that diverse teams performed equally as 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 nine in landscape models where $n = 10, k = 3, l = 9$, smoothness factors ranging from 1 to 6, and trust factors 0, 0.33, 0.5, and 1 (see (Huang 2024) for more details).

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

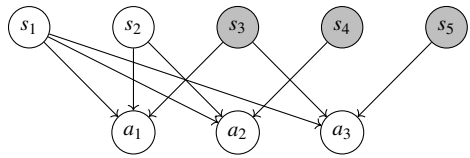


Figure 1. A stylized evidential sources model consisting of five sources s_1–s_5 and three agents a_1–a_3. The coloring of the sources represents a stylized model run, where the evidence produced by the white sources (resp., gray sources) was (inferentially) accurate (resp., inaccurate). Observe that in this particular model run: (a) agents 1 and 2 have a correct belief, while agent 3's belief is incorrect; (b) the opinion-based dynamics leads to a correct collective decision (because the majority of the agents have a correct belief); (c) the evidence-based dynamics leads to an incorrect collective decision (because the majority of the sources accessed by the team produced inaccurate evidence); and (d) the boundedly rational evidence-based dynamics leads to a correct collective decision (because the accurate evidences are accessed five times (including double counting) while the inaccurate evidences are only accessed four times).

3. The evidential sources model

In this section I present the new evidential sources framework (section 3.1), discuss and operationalize the concept of diversity at the heart of my study (section 3.2), and discuss some numerical examples that lead to some preliminary findings (section 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. While 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 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 section 4.1 and will later use it to explore the implications for epistemic deliberative democracy in section 5.

3.1. The model

My new modeling 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, a reputable newspaper is (presumably) 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 first-year philosophy student.

Models consist of sources, agents, and teams. Figure 1 presents a stylized example with three agents and five sources, where edges represent the sources that are accessed by the agents.[10] Let us discuss these in turn. *Sources* can be more or less

[10] One could view Figure 1 as a causal network representing the dependencies (Dietrich and Spiekermann 2013a,b), except that it should be expanded with a node representing the state of the world, which can take two values, representing the correct option, and which directly causally affects all source nodes.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

reliable and some could even be systematically unreliable. 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 captures which option it supports. We assume that (the valences of the evidences generated by) the sources are probabilistically independent.

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 modeled 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, 2, and 4. 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 majority rule. An agent's *individual score* represents their level of *expertise*; it is the probability that the agent succeeds in selecting the correct option.[11] An agent's score can be calculated from the reliability of her sources by going over all possible configurations of valences of the evidences of her sources. This opinion formation 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.[12]

In my model, higher levels of expertise are correlated with repeated success.[13] Let me explain. The individual score is the probability that an agent succeeds in selecting the correct option, averaging over all those configurations of valences of the evidences of the agent's sources that lead to a correct opinion (weighted by the

[11] This belief formation mechanism relates to the traditional voting literature on heterogeneous, independent voters, where voters would correspond to the sources in the agent's classification heuristic and the majority competence would correspond to the agent's score.

[12] One possible extension could be that agents have (subjective or imperfect) assessments of the sources' relative reliability (see Lehrer and Wagner 1981; Grofman et al. 1983; Klein and Sprenger 2015). In Duijf (2024), I developed a model (similar to the evidential sources model) where the principal agent has imperfect capacities to evaluate the sources' reliability and the evidences' (inferential) accuracy, and use it to explicate the conditions that must obtain if open-mindedness is to be epistemically valuable.

[13] In contrast, Holman et al. (2018) and Grim et al. (2019) convincingly argue that experts in Hong and Page's original landscape model are not repeatedly successful.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

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.[14] Hence, higher individual scores represent higher likelihoods of repeated success.

Teams are modeled 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 opinion.
(2) Evidence-based dynamics: Agents communicate all their evidences and the team's collective decision follows the majority evidence.
(3) Boundedly rational evidence-based dynamics: Agents communicate all their evidences and the team's collective decision follows the majority evidence—possibly with double counting.

Figure 1 illustrates these different mechanisms in a stylized example. I will discuss these different deliberative mechanisms in more detail in section 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).

3.2. Operationalizing diversity

Cognitive diversity refers to the variety of perspectives, interpretations, heuristics, and predictive models people use to form beliefs and make choices.[15] In my model, agents form their beliefs and make decisions by applying the majority rule on the (valences of the) evidences generated by the sources they have access to. My notion of evidences is very broad and includes not only empirical facts but also opinions, observations, arguments, intuitions, and personal experiences. Therefore, an agent's classification heuristic, which is represented by the sources she has access to, can be interpreted as her perspective (how she views the problem) or her decision heuristic

[14] Goldman (2018, 5) concurs that an expert with respect to a given domain "possesses more and/or better evidence pertaining to propositions in [that domain] than most people in the relevant comparison class."

[15] Landemore (2013) writes: "What matters is a more fundamental cognitive diversity, defined as the internal, psychological property that determines how each individual sees the world, interprets its problems, and makes predictions in it."

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

(how she makes her decisions). Accordingly, I operationalize diversity as a lack of overlap in the agents' classification heuristics (i.e., a lack of overlap in the agents' evidential sources).[16] For example, in Figure 1, agents 1 and 2 are less diverse than agents 2 and 3. In general, the more diverse two agents' classification heuristics are, the lower the correlation between their decisions will be.

More technically, the *diversity* of two heuristics A and B, $\Delta(A, B)$, is defined by the proportion of sources that are in only one of them:[17]

$$\Delta(A, B) = \frac{1}{2}\left(\frac{|A \backslash B|}{|A|} + \frac{|B \backslash A|}{|B|}\right).$$

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

$$\Delta(\mathscr{T}) = \gamma \cdot \sum_{\substack{A,B \in \mathscr{T} \\ A \neq B}} \Delta(A, B),$$

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

3.3. Examples and preliminary findings

In this section, we discuss some numerical examples and draw out some preliminary findings and observations. First, to investigate the diversity–expertise trade-off, we must consider situations where the heuristics are not equally good. Hence, we must consider evidential sources models where the sources are not equally reliable (otherwise, all heuristics would be equally good).

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

[16] I deliberately try to stay as close as possible to the Hong–Page tradition in this regard. In the Hong–Page tradition, search heuristics represent ways of generating potential solutions to problems. These heuristics are modeled by finite strings of numbers, where these numbers represent the potential solutions to check. Diversity is modeled as the (lack of) overlap in such strings of numbers, i.e., the (lack of) overlap in the potential solutions they check. Moreover, Dietrich and Spiekermann (2025, 607), working within a framework similar to mine, write: "Diversity is key to successful deliberation. It can be construed as heterogeneity in sources, i.e., dissimilarity between the source sets (the arguments, the empirical knowledge, etc.) of different persons."

[17] It may be helpful to note that (i) $\Delta(A, B) = \Delta(B, A)$, (ii) $0 \leq \Delta(A, B) \leq 1$, (iii) $\Delta(A, B) = 1$ iff $A \cap B = \emptyset$, and (iv) $\Delta(A, B) = 0$ iff $A = B$.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

diversity–expertise trade-off, 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 three. 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 trade-off.

4. Simulation studies

In this section, I discuss and motivate the parameter choices in my simulation studies (section 4.1) and then present the results of my simulation studies concerning opinion-based (section 4.2), evidence-based (section 4.3), and boundedly rational evidence-based (section 4.4) dynamics. I discuss various robustness checks in section 4.5.

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.[18]

[18] In online supplement C, I report on several robustness checks: teams of size 5 and 7; heuristics of size 7 and 9; heterogeneous heuristics of size 5, 7, or 9; 21 available sources; and equidistant reliability distributions with a smaller and wider range.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

First, following the existing literature on landscape models, I decided to focus on teams of size nine—although this decision is, admittedly, rather arbitrary.[19]

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 three and five pieces of information (Cowan 2001, 2010), I decided to only consider heuristics of size 5.[20,21] After all, the size of an agent's heuristic represents the number of evidences used in their individual belief-formation and decision-making.

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).[22] The set of all possible heuristics (or agents) in a given parameter setting is determined by the number of available sources n and the size of the heuristics k; their total number is given by $\binom{n}{k}$. Given my focus on heuristics of size 5, I considered models with 13 or 17 sources, which correspond to 1,287 $(= \binom{13}{5})$ and 6,188 $(= \binom{17}{5})$ possible heuristics, respectively.

Fourth, my choice regarding the source reliability distributions is mainly guided by simplicity and tractability. I focused on, so-called, equidistant source reliability distributions with a fixed range of 0.2. These reliability distributions are represented by their mean. For example, mean 0.7 corresponds to an equidistant reliability distribution on the interval $[0.6, 0.8]$.[23]

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 agent 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

[19] 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 (with results for sizes 3 and 6 in their appendix).

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

[21] One might object that experts (or even lay people) often have access to more than five sources. Let me briefly mention three responses: (1) In my view, the parameter choice is compatible with an interpretation that agents have access to more than five sources, but only five of those sources have a significant impact on their decision-making. (2) I believe the adequacy of this parameter choice depends more on the 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 (see Figure 2). (3) In online supplement C, I considered larger heuristic sizes, which made expertise more valuable in all circumstances. I thank Felix Kopecky and Max Noichl for pressing me on this point.

[22] 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.

[23] In general, for n sources with mean reliability $\bar{p}$, the equidistant distribution would correspond to the set of source reliabilities given by $\{\bar{p} - 0.1 + \frac{k}{n} \cdot 0.2 | k \in \mathbb{N}, k \leq n\}$. For example, for 21 sources and mean reliability 0.55, the corresponding set would be {0.45, 0.46, 0.47, …, 0.64, 0.65}.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

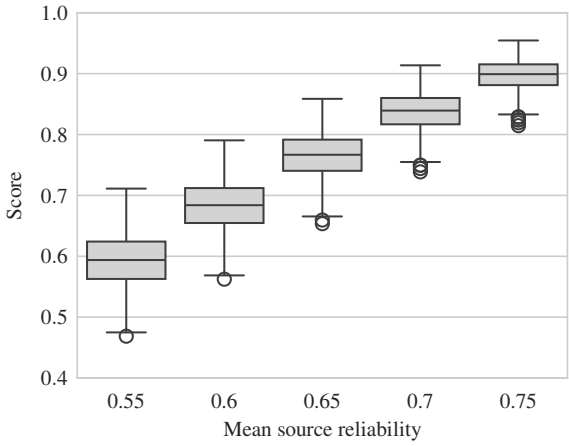


Figure 2. Individual scores.

the highest individual scores. The *diverse* teams were generated by sequentially adding members using a greedy search procedure.[24] Subsequently, I calculated the teams' scores for each of the three deliberative mechanisms by going over all possible configurations of valences of the evidences of the sources. For a given deliberative mechanism, the team's score is given by the sum of the probabilities associated with those configurations that led to a correct team decision.

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.[25] 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 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 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 lay people; 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 agents differs. From

[24] Roughly stated, at each step of this search procedure, a new heuristic that maximizes diversity is randomly selected. Alternatively, one could consider all possible teams and pick 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.

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

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

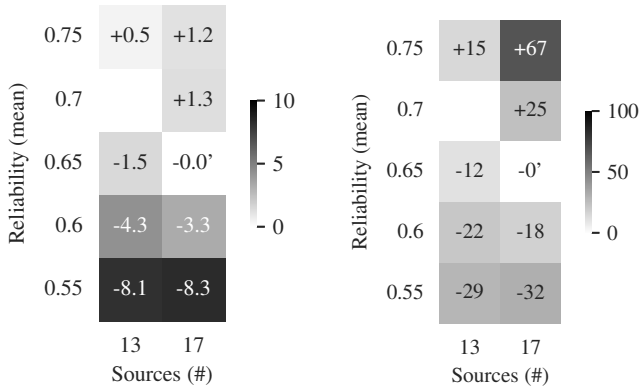


Figure 3. The diversity–expertise trade-off for opinion-based dynamics.

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.

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—and then follow the majority opinion. 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). This two-step mechanism relates to the voting literature on heterogeneous, correlated voters (Ladha 1992, 1993, 1995; Berg 1993; Pivato 2017) and on indirect majority systems (Boland et al. 1989; Berg 1997).[26]

Figure 3 presents the results of my study, showing the performance difference between diverse teams and expert teams, that is, the difference between the median score of 10,000 diverse teams and the expert team's score.[27] It is instructive to

[26] The voting literature on correlated voters focuses on providing conditions when the majority competence converges to 1 as the group size goes to infinity, or when the majority competence exceeds the mean competence (Ladha 1992, 1993, 1995; Berg 1993; Pivato 2017). The existing work on indirect majority systems focuses on situations where voters are divided into non-overlapping sets with homogeneous competence levels (Boland et al. 1989; Berg 1997). In contrast, my opinion-based dynamics could be viewed as an indirect majority system where voters (i.e., sources) are divided into (possibly) overlapping sets with heterogeneous reliability levels. Lastly, another influential model is the opinion-leader model (Boland et al. 1989, Model 1). Although the opinion-leader model enables some theoretical results (Boland et al. 1989; Karge 2025), it can only model a simple kind of dependency and cannot capture the complex dependency structures reminiscent of my model of opinion-based dynamics.

[27] Let me mention some conventions: Positive numbers indicate parameter settings where diversity trumped expertise; negative numbers represent those where expertise trumped diversity. Numbers represent the absolute performance difference (left) and relative difference (right) in percentages. I performed (one-sample) Wilcoxon tests. Results are masked if (and only if) $p \geq 0.001$. Following guidelines by Cohen (1988), effect sizes were calculated as $r = Z/\sqrt{N}$ and interpreted as follows: a single apostrophe (') indicates that the effect size was small ($r < 0.1$); two apostrophes (") indicate that the

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

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%.

The results clearly indicate that there is no universal winner: diverse teams outperform expert teams in some contexts, and vice versa. The trade-off 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 when 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 section 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 (2019) 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 virtually all simulation results were statistically significant and had a large effect size, 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 trade-off concerns the trade-off between selecting for ability or diversity. Let us start with ability. One would expect that higher individual performance differences favor the expert team. Figure 4 presents the *ability gap*: the median difference in the average individual scores between members of the expert team and the diverse team, showing both absolute and relative performance differences. Observe that the absolute ability gap

effect size was medium ($0.1 \leq r < 0.3$); no apostrophe indicates that the effect size was large ($0.3 \leq r$). Online supplement A contains a full report, also including ratios and 95% confidence intervals.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

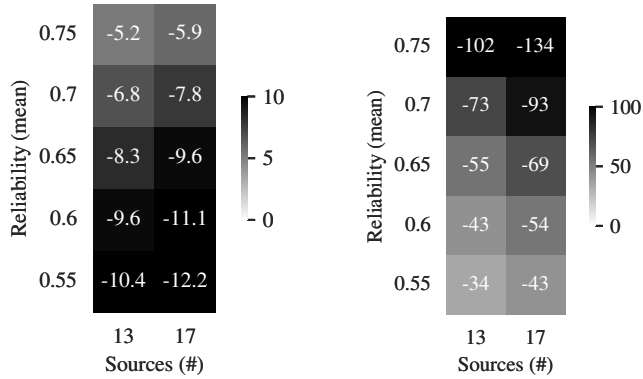


Figure 4. The ability gap.

narrows when the source reliabilities increase, while the relative ability gap widens. Hence, the absolute measure provides the best insight into the diversity–expertise trade-off.

We proceed with diversity. The difference in diversity between the expert team and diverse teams was 0.41 and 0.51 for scenarios with 13 and 17 sources, respectively. For example, when 13 sources were 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 five, this meant that, on average, members of the expert team had two out of five sources more in common compared with members of the diverse team. To grasp the extent of the correlation between experts, we can show that the *agreement rate between experts is high*: the likelihood that two experts agree varies between 75% and 94% and increases when sources become more reliable (see online supplement B).

The fairly high degree of correlation 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 online supplement B). Moreover, in all scenarios where the expert *team* trumped diverse teams, the highest-scoring *individual* also trumped diverse teams.

In conclusion, the ability gap and degree of correlation among experts jointly explain the diversity–expertise trade-off 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 ability gap 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 ability gap 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

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

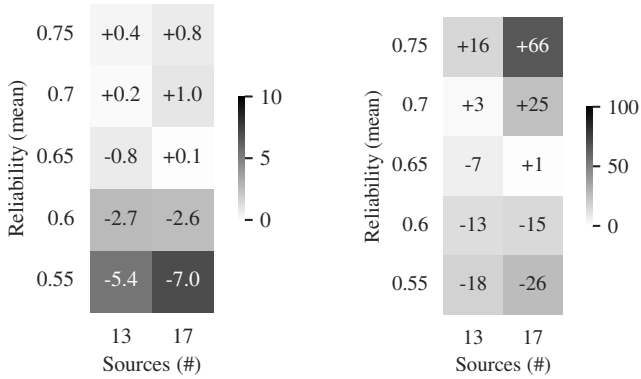


Figure 5. The diversity–expertise trade-off for evidence-based dynamics.

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 trade-off. This mechanism is related to the voting literature on heterogeneous, independent voters (Grofman 1975, 1978; Margolis 1976; Feld and Grofman 1984; Owen et al. 1989), where voters would correspond to the sources accessed by a given team in my model and the majority competence would correspond to the team's score for the evidence-based dynamics.

In any case, although we previously observed that expertise often trumped diversity for the opinion-based dynamics, it is plausible to think that the *evidence-based* dynamics favors 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 trade-off 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 favor 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 the evidence-based dynamics had an approximately similar relative effect on the performance of expert and diverse teams, while having a disparate absolute effect.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

What explains this diversity–expertise trade-off? 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 nine sources. So, the diverse team covered four and eight 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 expertise trumped diversity 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 trade-off for evidence-based dynamics is mainly explained by the sources' reliabilities.

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; Grofman et al. 1983).

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* counting. On the other hand, it seems plausible that skewed counting is only suboptimal when it leads to overcounting evidence from certain sources *regardless of their reliability*. Given that the expert team is overcounting evidence from the most reliable sources, overcounting appears to track the fact that certain sources are more reliable. By contrast, since diverse teams count evidence from all sources roughly equally, counting appears to be divorced from the sources' reliability. Hence, diverse teams appear to be more vulnerable to *arbitrary* counting.

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 trade-off are virtually always irrespective of the deliberative mechanism—the only exception is the scenario with 17 sources and a mean reliability of 65%.

Second, in scenarios where diversity trumped expertise, the boundedly rational dynamics further expanded the performance advantage of diverse teams (compared

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

Table 1. Source weights for the expert teams

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

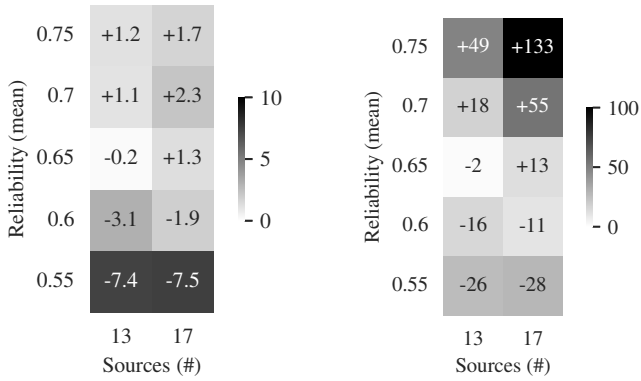


Figure 6. The diversity–expertise trade-off for boundedly rational evidence-based dynamics.

with the other deliberative mechanisms). In contrast, in scenarios where some sources were anti-reliable, the boundedly rational dynamics increased the deficit of diverse teams compared with the evidence-based dynamics. Hence, the risk of arbitrary counting in diverse teams appears to be outweighed by the risk of skewed counting 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 online supplement B). Nonetheless, whenever the expert team trumped diverse teams, the best-performing *individual* also trumped diverse teams. So, the qualitative results concerning the diversity–expertise trade-off appear to hinge on the best-performing individual.

In comparison, the diverse team counted all sources roughly equally. 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 online supplement B). This confirms the suspicion that counting 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 trade-off 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

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

accesses the top nine sources, in this case, the expert team acted similar to an individual who accesses the top five sources.

4.5. Robustness

The central result of the simulation study is two-fold.

(1) Diversity trumped expertise when sources were (at least weakly) reliable, when the problem was easy, and/or when the agents were already quite expert.
(2) Expertise trumped diversity when some sources were anti- or unreliable, when the problem was hard, and/or when the pool of agents contained novices.

I performed various robustness checks and considered smaller teams, larger heuristics, heterogeneous heuristic sizes, more sources, and source reliability distributions with narrower and wider ranges (details can be found in online supplement C). Overall, the robustness analysis shows that the result concerning when expertise trumped diversity is robust, but the result concerning when diversity trumped expertise is *not robust*. So, the circumstances where diversity trumps expertise might be even more constrained. In general, the robustness checks indicate that expertise is more valuable when the heuristic sizes are larger, heuristic sizes are diverse, there are fewer sources available, and/or the source reliabilities vary more.

5. Epistemic deliberative democracy

Epistemic democrats analyze the epistemic value of democracy and (typically) justify democracy based on its epistemic credentials (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). 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* justified.

These epistemic justifications of democracy are threatened by citizens' staggering lack of political knowledge (Somin 2013; Achen and Bartels 2017). But, despite their ignorance of political facts, citizens can successfully use informational cues and shortcuts in their political decision-making (Lupia 2006; Goren 2013). Moreover, comparisons between uninformed citizens and their informed counterparts show that uninformed citizens fall between a coin toss and the informed citizen (Bartels 1996; Lau and Redlawsk 1997, 2006). Hence, under the plausible assumption that informed citizens are better than chance level, the empirical evidence indicates that (uninformed) citizens are minimally competent in the sense that they are better than chance level at determining which political option is best. Although minimal competence levels may suffice to justify democratic *aggregation* methods (e.g., majority voting) on the basis of Condorcet-like jury theorems (Goodin and

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

Spiekermann 2018; *pace* Caplan 2007), it is an open question whether it suffices for *deliberative* democracy.

To determine the epistemic value of deliberative democracy, I will assume that deliberative democracy is operationalized in terms of small, diverse, deliberative citizen assemblies. Moreover, I will assume, like epistemic democrats, that there exists some procedure-independent measure of success. Do small, diverse, deliberative citizen assemblies perform better than expert teams on this epistemic score? According to my simulation study, this hinges on whether the evidential sources used by citizens are (at least weakly) reliable, whether political problems are easy, and/or whether the pool of citizens is already quite expert in the political domain. If so, diverse teams of citizens can outperform expert teams. If not, expert teams outperform diverse teams of citizens. Hence, 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 obtain a reasonable guestimate of citizens' competence. Experiments by Lau and Redlawsk (2006, 79) indicated that: "Seventy percent of the subjects (206 out of 293) would not have changed their vote after learning everything there was to learn about the two candidates."[28] Without going into the details of their experimental design and methodology, it is important to note that (hypothetical) informed voters are still not perfect. Although it is hard to predict how well these findings generalize, it seems reasonable to assume that citizens' competence, on average, falls between a coin toss and 70%. Now, Figure 2 illustrates that this corresponds to evidential sources models with mean source reliability of 55% and 60% in my simulation study.[29] Furthermore, my simulation study robustly showed that expert teams trumped diverse teams in these circumstances. One could even argue that my simulation study paints an overly pessimistic picture of expert teams, because experts would be even more competent than the best agents in many parameter settings. In any case, I conclude that, from a purely epistemic perspective, when selecting for small, deliberative assemblies, one should prioritize expertise over diversity.[30]

I should add one important caveat. The previous argument assumes that the deliberative assemblies are made up of fairly uninformed citizens. Instead, deliberative democrats could propose more elaborate procedures that ensure that the participants in the deliberative citizen assemblies gain relevant knowledge and competence (e.g., Fishkin 2009); for example, they may receive balanced briefing materials prior to deliberation. Overall, my simulation study suggests that the epistemically best teams are diverse teams consisting of agents who are already quite

[28] Other important examples that estimate and simulate so-called enlightened preferences or fully informed voters include those by Bartels (1996) and Althaus (2003).

[29] It may be helpful to add that, in view of Figure 2, parameter settings where the mean source reliability is 65% correspond to pools of agents with a fairly high average individual score of 77% (first and third quartile scores of 74% and 79%, respectively); those where the mean source reliability is 60% correspond to an average individual score of 68% (first and third quartile scores of 65% and 71%, respectively).

[30] Of course, these epistemic considerations may be overruled by other non-epistemic considerations, such as political or moral considerations.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

expert. Hence, one might conjecture that diverse, deliberative assemblies of fully informed citizens outperform teams of experts.[31]

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 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 analyzed 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 robustly 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. But, since the latter finding was not robust across reasonable variations in the model's parameters, the conditions when diversity trumps expertise may be more constrained.

What is the implication for epistemic deliberative democracy? Although citizens' knowledge of political facts is staggeringly low, citizens appear to be minimally competent in the political domain (for instance, by relying on cues or shortcuts). While minimal competence may suffice for epistemic justifications of aggregative forms of democracy, I argued, on the basis of my simulation study, that the epistemic credentials of deliberative democracy are questionable. After all, in those parameter settings that were compatible with reasonable estimates of citizens' competence, expert teams trumped diverse teams.

All in all, my simulation study demonstrates that diversity and expertise are both important—diversity addresses the risk of suboptimal coverage, expertise that of unreliable evidence. My results suggest that optimizing the team composition will require selecting for diversity with some epistemic guardrails. So, to improve collective knowledge acquisition we must strike the right balance between expertise and diversity.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/psa.2026.10200. All scripts necessary to reproduce the results are available on Zenodo (Duijf,

[31] Given my simulation study, this is a mere conjecture for two main reasons: first, the result concerning when diversity trumped expertise was not robust (see section 4.5), and second, my simulation study was not intended to capture such elaborate procedures.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

2025) and on GitHub (https://github.com/HeinDuijf/DiversityAbility), including a walkthrough of the model on a GitHub page.

Acknowledgments. I am grateful to Daniel Cohnitz, Janneke van Dis, Dominik Klein, Max Noichl, Uwe Peters, and Kai Spiekermann for their thoughtful feedback on earlier drafts of this paper. I would like to thank the Centre for Digital Humanities at Utrecht University, in particular Jonathan de Bruin, Kirsten Schutter, and Jelle Treep, for their invaluable support. I also thank the audiences at presentations in Amsterdam, Bochum, Düsseldorf, Groningen, Luxembourg, Paris, and York, whose valuable questions and suggestions helped me improve this work in many ways.

Funding statement. None to declare. Open access funding provided by Utrecht University.

Declarations. None to declare.

References

Achen, Christopher H., and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400888740

Althaus, Scott L. 2003. *Collective Preferences in Democratic Politics: Opinion Surveys and the Will of the People*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511610042

Anderson, Elizabeth. 2011. "Democracy, Public Policy, and Lay Assessments of Scientific Testimony." *Episteme* 8 (2):144–64. https://doi.org/10.3366/epi.2011.0013

Armstrong, J. Scott. 2001. "Combining Forecasts." In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, edited by J. Scott Armstrong. Boston, MA: Springer, pp. 417–39. https://doi.org/10.1007/978-0-306-47630-3_19

Axelrod, Robert. 1997. "The Dissemination of Culture: A Model with Local Convergence and Global Polarization." *Journal of Conflict Resolution* 41 (2):203–26. https://doi.org/10.1177/0022002797041002001

Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40 (1):194–230. https://doi.org/10.2307/2111700

Batchelor, Roy, and Pami Dua. 1995. "Forecaster Diversity and the Benefits of Combining Forecasts." *Management Science* 41 (1):68–75. https://doi.org/10.1287/mnsc.41.1.68

Berg, Sven. 1993. "Condorcet's Jury Theorem, Dependency Among Jurors." *Social Choice and Welfare* 10 (1):87–95. https://doi.org/10.1007/BF00187435

Berg, Sven. 1997. "Indirect Voting Systems: Banzhaf Numbers, Majority Functions and Collective Competence." *European Journal of Political Economy* 13 (3):557–73. https://doi.org/10.1016/S0176-2680(97)00014-1

Boland, Philip J., Frank Proschan, and Y. L. Tong. 1989. "Modelling Dependence in Simple and Indirect Majority Systems." *Journal of Applied Probability* 26 (1):81–88. https://doi.org/10.2307/3214318

Caplan, Bryan. 2007. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400828821

Cohen, Joshua. 1986. "An Epistemic Conception of Democracy." *Ethics* 97 (1):26–38. https://doi.org/10.1086/292815

Cohen, Joshua. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. New York: Routledge. https://doi.org/10.4324/9780203771587

Cowan, Nelson. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24 (1):87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, Nelson. 2010. "The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why?" *Current Directions in Psychological Science* 19 (1):51–57. https://doi.org/10.1177/0963721409359277

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

Dechêne, Alice, Christoph Stahl, Jochim Hansen, and Michaela Wänke. 2010. "The Truth About the Truth: A Meta-analytic Review of the Truth Effect." *Personality and Social Psychology Review* 14 (2):238–57. https://doi.org/10.1177/1088868309352251

DeGroot, Morris H. 1974. "Reaching a Consensus." *Journal of the American Statistical Association* 69 (345):118–21. https://doi.org/10.2307/2285509
Dietrich, Franz, and Kai Spiekermann. 2013a. "Epistemic Democracy with Defensible Premises." *Economics & Philosophy* 29 (1):87–120. https://doi.org/10.1017/S0266267113000096
Dietrich, Franz, and Kai Spiekermann. 2013b. "Independent Opinions? On the Causal Foundations of Belief Formation and Jury Theorems." *Mind* 122 (487):655–85. https://doi.org/10.1093/mind/fzt074
Dietrich, Franz, and Kai Spiekermann. 2023. "Jury Theorems." In *The Stanford Encyclopedia of Philosophy*, edited by Edward M. Zalta and Uri Nodelman. Stanford, CA: Stanford University Press.
Dietrich, Franz, and Kai Spiekermann. 2025. "Deliberation and the Wisdom of Crowds." *Economic Theory* 79 (2):603–55. https://doi.org/10.1007/s00199-024-01595-4
Ding, Huihui, and Marcus Pivato. 2021. "Deliberation and Epistemic Democracy." *Journal of Economic Behavior & Organization* 185:138–67. https://doi.org/10.1016/j.jebo.2021.02.020
Duijf, Hein. 2024. "When Should One Be Open-Minded?" *Philosophical Studies* 181 (5):1257–96. https://doi.org/10.1007/s11098-024-02137-3
Duijf, Hein. 2025. *Diversity and Ability in an Evidential Sources Framework. Zenodo.* https://doi.org/10.5281/zenodo.17899031
Estlund, David. 2008. *Democratic Authority: A Philosophical Framework*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400831548
Estlund, David, and Hélène Landemore. 2018. "The Epistemic Value of Democratic Deliberation." In *The Oxford Handbook of Deliberative Democracy*, edited by Andre Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark D. Warren. Oxford: Oxford University Press, pp. 112–31. https://doi.org/10.1093/oxfordhb/9780198747369.013.26
Feld, Scott L., and Bernard Grofman. 1984. "The Accuracy of Group Majority Decisions in Groups with Added Members." *Public Choice* 42 (3):273–85. https://doi.org/10.1007/BF00124946
Fishkin, James S. 2009, *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199604432.001.0001
Galton, Francis. 1907. "Vox Populi." *Nature* 75 (1949):450–51. https://doi.org/10.1038/075450a0
Genta, Benjamin Santos. 2024. "Formal Models and Justifications of Democracy." *Synthese* 204:138. https://doi.org/10.1007/s11229-024-04783-w
Goldman, Alvin I. 2001. "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research* 63 (1):85–110. https://doi.org/10.1111/j.1933-1592.2001.tb00093.x
Goldman, Alvin I. 2018. "Expertise." *Topoi* 37 (1):3–10. https://doi.org/10.1007/s11245-016-9410-3
Goodin, Robert E. and Kai Spiekermann. 2018. *An Epistemic Theory of Democracy*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198823452.001.0001
Goren, Paul. 2013. *On Voter Competence*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195396140.001.0001
Grim, Patrick, Daniel J. Singer, Aaron Bramson, Bennett Holman, Sean McGeehan, and William J. Berger. 2019. "Diversity, Ability, and Expertise in Epistemic Communities." *Philosophy of Science* 86 (1):98–123. https://doi.org/10.1086/701070
Grofman, Bernard. 1975. "A Comment on 'Democratic Theory: A Preliminary Mathematical Model'." *Public Choice* 21 (1):99–103. https://doi.org/10.1007/BF01705949
Grofman, Bernard. 1978. "Judgmental Competence of Individuals and Groups in a Dichotomous Choice Situation: Is a Majority of Heads Better than One?" *Journal of Mathematical Sociology* 6 (1):47–60. https://doi.org/10.1080/0022250X.1978.9989880
Grofman, Bernard, Guillermo Owen, and Scott L. Feld. 1983. "Thirteen Theorems in Search of the Truth." *Theory and Decision* 15 (3):261–78. https://doi.org/10.1007/BF00125672
Hardwig, John. 1985. "Epistemic Dependence." *The Journal of Philosophy* 82 (7):335–49. https://doi.org/10.2307/2026523
Hasher, Lynn, David Goldstein, and Thomas Toppino. 1977. "Frequency and the Conference of Referential Validity." *Journal of Verbal Learning and Verbal Behavior* 16 (1):107–12. https://doi.org/10.1016/S0022-5371(77)80012-1
Hegselmann, Rainer, and Ulrich Krause. 2002. "Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation." *Journal of Artificial Societies and Social Simulation* 5 (3):2.

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

Holman, Bennett, William J. Berger, Daniel J. Singer, Patrick Grim, and Aaron Bramson. 2018. "Diversity and Democracy: Agent-Based Modeling in Political Philosophy." *Historical Social Research/Historische Sozialforschung* 43 (1):259–84. https://doi.org/10.12759/hsr.43.2018.1.259-284

Hong, Lu, and Scott E. Page. 2004. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Academy of Sciences* 101 (46):16385–89. https://doi.org/10.1073/pnas.0403723101

Hong, Lu, and Scott E. Page. 2025. "The Range of Collective Accuracy for Binary Classifications Under Majority Rule." *Economic Theory* 79 (1):275–300. https://doi.org/10.1007/s00199-024-01570-z

Huang, Alice C. W. 2024. "Landscapes and Bandits: A Unified Model of Functional and Demographic Diversity." *Philosophy of Science* 91 (3):579–94. https://doi.org/10.1017/psa.2023.169

Karge, Jonas. 2025. "When Diversity in Voting Trumps Ability." *Episteme.* https://doi.org/10.1017/epi.2025.10042

Keuschnigg, Marc, and Christian Ganser. 2017. "Crowd Wisdom Relies on Agents' Ability in Small Groups with a Voting Aggregation Rule." *Management Science* 63 (3):818–28. https://doi.org/10.1287/mnsc.2015.2364

Klein, Dominik, and Jan Sprenger. 2015. "Modelling Individual Expertise in Group Judgements." *Economics & Philosophy* 31 (1):3–25. https://doi.org/10.1017/S0266267114000388

Kuehn, Daniel. 2017. "Diversity, Ability, and Democracy: A Note on Thompson's Challenge to Hong and Page." *Critical Review* 29 (1):72–87. https://doi.org/10.1080/08913811.2017.1288455

Ladha, Krishna K. 1992. "The Condorcet Jury Theorem, Free Speech, and Correlated Votes.: *American Journal of Political Science* 36 (3):617–34. https://doi.org/10.2307/2111584

Ladha, Krishna K. 1993. "Condorcet's Jury Theorem in Light of de Finetti's Theorem." *Social Choice and Welfare* 10 (1):69–85. https://doi.org/10.1007/BF00187434

Ladha, Krishna K. 1995. Information Pooling Through Majority-Rule Voting: Condorcet's Jury Theorem with Correlated Votes." *Journal of Economic Behavior & Organization* 26 (3):353–72. https://doi.org/10.1016/0167-2681(94)00068-P

Lamberson, P. J. and Scott E. Page. 2012. "Optimal Forecasting Groups." *Management Science* 58 (4):805–10. https://doi.org/10.1287/mnsc.1110.1441

Landemore, Hélène. 2013. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400845538

Landemore, Hélène, and Scott E. Page. 2015. "Deliberation and Disagreement: Problem Solving, Prediction, and Positive Dissensus." *Politics, Philosophy & Economics* 14 (3):229–54. https://doi.org/10.1177/1470594X14544284

Lau, Richard R., and David P. Redlawsk. 1997. "Voting Correctly." *The American Political Science Review* 91 (3):585–98. https://doi.org/10.2307/2952076

Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing during Election Campaigns. New ork: Cambridge University Press.* https://doi.org/10.1017/cbo9780511791048

Lehrer, Keith, and Carl Wagner. 1981. *Rational Consensus in Science and Society*. Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-8520-9

Liggett, Thomas M. 1985. *Interacting Particle Systems*. Berlin: Springer. https://doi.org/10.1007/978-1-4613-8542-4

Lupia, Arthur. 2006. "How Elitism Undermines the Study of Voter Competence." *Critical Review* 18 (1–3):217–32. https://doi.org/10.1080/08913810608443658

Margolis, Howard. 1976. "A Note on Incompetence." *Public Choice* 26 (1):119–27. https://doi.org/10.1007/BF01725800

Mäs, Michael, and Andreas Flache. 2013. "Differentiation Without Distancing. Explaining Bi-polarization of Opinions Without Negative Influence." *PLOS One* 8 (11):e74516. https://doi.org/10.1371/journal.pone.0074516

Nowak, Szamrej, and Latanà Nowak, A., J. Szamrej, and B. Latané. 1990. "From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact." *Psychological Review* 97 (3):362–76. https://doi.org/10.1037/0033-295X.97.3.362

Owen, Guillermo, Bernard Grofman, and Scott L. Feld. 1989. "Proving a Distribution-Free Generalization of the Condorcet Jury Theorem." *Mathematical Social Sciences* 17 (1):1–16. https://doi.org/10.1016/0165-4896(89)90012-7

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press

Page, Scott. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j.ctt7sp9c
Pivato, Marcus. 2017. "Epistemic Democracy with Correlated Voters." *Journal of Mathematical Economics* 72:51–69. https://doi.org/10.1016/j.jmateco.2017.06.001
Reijula, Samuli, and Jaakko Kuorikoski. 2021. "The Diversity–Ability Trade-Off in Scientific Problem Solving." *Philosophy of Science* 88 (5):894–905. https://doi.org/10.1086/714938
Rolin, Kristina H. 2021. "Objectivity, Trust and Social Responsibility." *Synthese* 199 (1):513–33. https://doi.org/10.1007/s11229-020-02669-1
Sakai, Ryota. 2020. "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. https://doi.org/10.1177/0048393120917635
Singer, Daniel J. 2019. "Diversity, not Randomness, Trumps Ability." *Philosophy of Science* 86 (1):178–91. https://doi.org/10.1086/701074
Somin, Ilya. 2013. *Democracy and Political Ignorance*. Stanford, CA: Stanford University Press.
Steel, Daniel, Sina Fazelpour, Bianca Crewe, and Kinley Gillette. 2021. "Information Elaboration and Epistemic Effects of Diversity." *Synthese* 198 (2):1287–307. https://doi.org/10.1007/s11229-019-02108-w
Sulik, Justin, Bahador Bahrami, and Ophelia Deroy. 2022. "The Diversity Gap: When Diversity Matters for Knowledge." *Perspectives on Psychological Science* 17 (3):752–67. https://doi.org/10.1177/17456916211006070
Surowiecki, James. 2005. *The Wisdom of Crowds*. New York: Knopf Doubleday Publishing Group.
Thompson, Abigail. 2014. "Does Diversity Trump Ability?" *Notices of the AMS* 61 (9):1024–30. https://doi.org/10.1090/noti1163
Weymark, John A. 2015. "Cognitive Diversity, Binary Decisions, and Epistemic Democracy." *Episteme* 12 (4):497–511. https://doi.org/10.1017/epi.2015.34

Cite this article: Duijf, Hein. 2026. "Diversity and Expertise in Binary Classification Problems." *Philosophy of Science* 93 (3):631–654. https://doi.org/10.1017/psa.2026.10200

https://doi.org/10.1017/psa.2026.10200 Published online by Cambridge University Press