Philosophy of Science (2025), **92**, 40–58
doi:10.1017/psa.2024.29

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

ARTICLE

Inconsistent Belief Aggregation in Diverse and Polarised Groups

Felix Kopecky and Gregor Betz

DebateLab, Karlsruhe Institute of Technology, Karlsruhe, Germany
Corresponding author: Felix Kopecky; Email: f.kopecky@kit.edu

(Received 09 October 2023; revised 01 February 2024; accepted 13 June 2024; first published online 01 October 2024)

Abstract

How do opinion diversity and belief polarisation affect epistemic group decision-making, particularly if decisions must be made without delay and on the basis of permissive evidence? In an agent-based model, we track the consistency of group opinions aggregated through sentence-wise majority voting. Simulations on the model reveal that high opinion diversity, but not polarisation, incurs a significant inconsistency risk. These results indicate that epistemic group decisions based on permissive evidence can be particularly difficult for diverse groups. The results also improve our understanding of what can reasonably be expected of expert groups, and where expert advice might have limits.

1. Introduction

When citizens and their governments rely on expert groups for policy advice, should they favour a diversified group composition? A view into the literature on epistemic problem solving seems to suggest an affirmative answer: groups with diverse viewpoints cover a high proportion of the approaches to a problem (Hong and Page 2004; Grim et al. 2019), and carry alternatives when popular hypotheses are falsified (Pöyhönen 2017; Zollman 2010). Provided sufficient time to explore the evidence and hypotheses, these effects underpin the benefits of opinion diversity in epistemic group problem solving. But what if the group's decision-making is constrained by limited time and the available evidence permits multiple yet incompatible responses? Is diversity likewise beneficial when decisive pieces of evidence and the optimal response are only discovered well after the experts gave their recommendation?

A similar question arises for groups that polarise while being pressed for time or lacking evidence. Should citizens and policy makers avoid polarised expert groups in uncertain situations and under time pressure? Although belief polarisation might appear to hamper decision-making by inhibiting consensus, some research indicates that polarised groups can be capable epistemic problem solvers (Shi et al. 2019) – again, at least in the long term.

Accompanied by an agent-based model based on the theory of dialectical structures (Betz 2009), we investigate how likely groups with different degrees of opinion diversity and belief polarisation are to form consistent group opinions when they use sentence-wise majority voting as part of their decision-making process. Majority voting appears particularly suitable when decisions must be made without delay, because it is easy to implement, almost instantaneous, and widely known. This makes it a formidable "closure device" (Richardson 2002, 203) to obtain a representative group opinion quickly when consensus cannot be reached otherwise. But it is also true that group opinions obtained through majoritarian aggregation can be inconsistent even if, as in our groups, all members hold consistent views (List and Pettit 2002).

In simulations on our model, we observe inconsistent majoritarian aggregation predominantly in highly diverse groups, but not in highly polarised or moderately diverse groups. This effect holds as long as the evidence is epistemically permissive, that is, several distinct and disagreeing belief systems are equally justifiable in light of the presented evidence.

Although we do not model the decision-making process of any particular group, we take the observed inconsistency prevalence as a useful indicator of problem difficulty. As alternatives to majority voting are likely to be more demanding, groups that seek consistent beliefs face additional tasks in preparing other procedures when they cannot include majority voting in their decision-making. We conclude that epistemic group problem solving can become more difficult for diverse groups in epistemically permissive situations – first, because they are more likely to require reflection on the aggregation procedure, and second, because inconsistencies are not automatically avoided by giving our agents more evidence, as long as this evidence remains permissive. Calls to increase opinion diversity in expert groups are well motivated by a presumed legitimacy boost, but the results from our simulations indicate that advice from diverse groups might have limits when decisions must be made without delay and on the basis of permissive evidence.

This paper begins with a description of the foundational concepts and our agent-based model in section 2. There we describe majoritarian belief aggregation and the inconsistent group beliefs that it is prone to. We also review the Gini–Simpson diversity index and a polarisation measure in this section, and describe our model to study epistemic decision problems. We then present results obtained from simulations on this model in section 3 and discuss possible consequences for expert advice in section 4. The conclusion (section 5) contains a brief summary of results and ramifications.

2. Belief aggregation, diversity and polarisation

2.1. A minimal example of inconsistent majoritarian belief aggregation

Imagine a group that is commissioned to form consistent beliefs about a set of propositions. Further suppose that the group is aware of a set of arguments expressing inferential relations between these propositions, and that all agents in the group agree that the presented arguments are pertinent and valid, and that no further arguments should be brought up at this time. Informed by the arguments, everyone in the group holds individually consistent but different views. We may

Table 1. Minimal example for an inconsistent sentence-wise majoritarian aggregation arising from the argument $(p_1 \wedge p_2) \rightarrow p_3$

Opinion of	p_1	p_2	$(p_1 \wedge p_2) \rightarrow p_3$	p_3
A1	T	F	T	F
A2	T	T	T	T
A3	F	T	T	F
Majority	T	T	T	F

assume that uncertainty surrounds the issue and that the arguments presented so far do not point to a uniquely optimal view but instead permit multiple justifiable responses. After discovering that they hold disagreeing views on the propositions and that there is no way to attain agreement for the moment, the agents decide to cast a vote on all propositions to form a representative group opinion. This vote, the members hope, should enable the group to make a recommendation, or at least support their further decision-making.

Even though they are in meta-agreement about the pertinence and validity of all arguments, the group must find that their sentence-wise majority vote is not guaranteed to yield a consistent group opinion (List and Pettit 2002). Table 1 illustrates a minimal example of such a case. Three agents hold beliefs that are individually consistent but are aggregated to group beliefs that are not. In the table, agreement on validity is expressed by the universal acceptance of the relation $(p_1 \wedge p_2) \rightarrow p_3$. But the agents differ in their beliefs otherwise. Agent A1 accepts p_1 but denies p_2, and so is able to reject the conclusion p_3. A2 accepts both premises and, by accepting the argument's validity, is obliged to accept the conclusion as well. Like A1, A3 rejects the conclusion but for a different reason: it accepts the premise p_2 while rejecting the premise p_1. The group opinion aggregated through sentence-wise majority voting is inconsistent: it rejects the conclusion while accepting the argument's validity and all of its premises.

Inconsistent belief aggregation is a problem for groups that issue recommendations to the public and policy makers. In section 2.2 we consider the difficulties associated with inconsistencies in group beliefs and paradigmatic scenarios in which they can plausibly arise. We then present our agent-based model that helps us understand whether diverse and polarised groups are more likely to encounter inconsistent majoritarian aggregation in decision problems that are more complex than the minimal example from this section.

2.2. The relevance of inconsistent aggregation in expert groups

Not all groups are equally affected by the risk of inconsistent aggregation, but it is a particular issue for expert groups when they convene to provide advice to policy makers and the public. We believe this is so for at least three reasons. Inconsistencies limit the utility of expert advice as it can involve recommendations that are mutually exclusive or defeat the purpose of other recommendations. Secondly, inconsistent opinions can question the very expertise of the group and its members. If an expert panel does not come up with consistent advice, maybe one should trust a different group with urgent questions? And thirdly, policies supported by inconsistent expert advice can lack *public justification*. The demand of public justification requires that policies should be justified in such a way that, in principle, anyone can understand and accept them (Vallier 2022). Not all justification for a policy is automatically lost in the case of inconsistent advice, but we find it plausible to assume that it will make the justification more complex.

The occurrence of an inconsistent majoritarian aggregation does not necessarily imply that the experts would relay inconsistent advice. But inconsistent aggregation is problematic even when experts become aware of it, because their original task of providing advice remains unresolved. We assume that inconsistent majoritarian aggregations would require groups to reflect on their aggregation procedure and that this would add to the difficulty of their decision problem. The occurrence of inconsistencies in belief aggregation thus indicates a particularly difficult instance of decision-making for expert groups.

This difficulty is exacerbated by the fact that ad hoc strategies to avoid inconsistencies are not particularly appealing to expert groups. Before turning to the question of how likely the risk of inconsistencies actually is, we briefly illustrate these "conclusion-driven" and "premise-driven" strategies (List and Pettit 2002, 93; Pettit 2001, 274) in the specific context of expert advice.

Expert groups are sometimes queried for an isolated proposition rather than a comprehensive and consistent set of recommendations. Policy makers might be interested solely whether to implement a particular policy or not. Following the conclusion-driven strategy, our experts would vote on a single proposition only and announce that outcome to the public. A recommendation on an isolated policy does not come with an inconsistency risk, but the problem is merely delayed. Inconsistencies could still emerge if the policy makers (or a curious subset of it) respond with critical questions regarding the recommended conclusion. The experts would then be expected to reply with reasons that are consistent and supported by a majority to back up their recommendation. The problem may also re-emerge even if critical questions are never raised. It is not at all unlikely that the expert group is asked, now or in the future, to issue further recommendations on other policies. If these policies are inferentially related to the first, inconsistencies might still arise, at which point all of their judgements could be doubted (an issue discussed by Pettit 2001, 279–80). And it might also turn out that the supposedly isolated policy issue is not that isolated but involves decisions on several inferentially related propositions.

The expert group could also pursue the premise-driven approach. They would vote on the premises only and determine their collective view of the conclusion by following the argument where it leads them. While this strategy would ensure consistency, it would lead to an unappealing outcome as well. Consider that the group would pursue this strategy in light of table 1. They would then accept the conclusion as there is majority support for all premises. But a majority of experts denies the conclusion! The resulting verdict would not be reflective of the expert opinions in the group.

This scenario of an unacceptable condition (inconsistent beliefs) where each available remedy is problematic (prioritising the premises or the conclusion) leads to the *discursive dilemma* (Pettit 2001, 274). Its occurrence is quite serious in theory and, as previous research shows, not at all improbable under plausible assumptions (List 2005).

If inconsistent aggregation can become problematic for expert groups, what are the scenarios in which it could arise? For the purpose of this study, we wish to differentiate three non-exhaustive but paradigmatic scenarios of majoritarian aggregation from individually consistent opinions on inferentially connected propositions:

(1) Someone external to the group polls the group members and aggregates the received opinions based on majority. Examples include parliamentary committees or research done by journalists. In this scenario we call *expert poll*, the experts do not communicate with each other for the purpose of aggregation.

(2) A group of experts meet under considerable time constraints to aggregate a recommendation. Although the experts are aware of each other's opinion and share a pool of available evidence, there is little time to evaluate novel evidence, and the experts' opinions are not changed in the meeting. The group decides to cast a majority vote as part of their decision-making. We label this scenario an *expert meeting*.

(3) A group of experts meet in conditions that allow them to review novel evidence and engage in prolonged discussions in which at least some change their views. We call this scenario *expert deliberation*. As uncertainties and disagreement remain, the experts still opt for voting.

These three scenarios represent relevant but exceptional circumstances for expert decision-making. In normal conditions, experts can expect to have sufficient time for deliberation and evidence accumulation to reduce uncertainty, and the public and policy makers are often content with receiving a diversity of views rather than a single consistent one. We will keep this in mind when discussing our results in section 4.

Our goal for the present paper is to understand whether groups in these paradigmatic scenarios are more or less likely to face inconsistent majoritarian aggregation depending on how diverse and polarised the opinions of their members are. To this end, we first review diversity and polarisation measures to identify populations with high diversity and polarisation (section 2.3), and then describe algorithms to synthesise agent samples in epistemic decision problems with varying numbers of arguments and specific levels of diversity and polarisation (section 2.4).

2.3. Measures of opinion diversity and belief polarisation

We begin this section with a review of the Gini–Simpson index, a quantified diversity measure, and then turn to statistical dispersion, a measure of belief polarisation. Beyond being the foundation for our analysis in section 3, an inspection of these measures also reveals how diverse and polarised groups differ. We turn to this at the end of this section.

In ecology, an ecosystem sample can be measured for diversity by calculating the frequencies of all species in the sample. The collected frequencies indicate how likely it is that an individual would encounter an individual of a different species. This is the core idea behind measuring diversity with the Gini–Simpson index (Tuomisto 2010, 856; Page 2011, chapter 2), and we find an analogous measure to be informative about opinion diversity as well.

To illustrate how the index works, suppose you mingle with the participants of an ethics conference. What is the probability that you will encounter a Kantian, a consequentialist, or a virtue ethicist? If the chance is about equal across all three groups, the conference crowd would be maximally diverse – relative to its members' opinions on moral theory. This kind of quantitative diversity analysis depends on a classification of individuals into types, just like establishing ecosystem diversity depends on knowledge about the individuals' species.

The Gini–Simpson index is related to other diversity indexes such as the Shannon index (Tuomisto 2010). It is an adequate measure for our model because it is refined on smaller populations (Tuomisto 2010, 854), an effect we can confirm for our populations of less than 100 individuals. We adjust the definition from Tuomisto (2010, 856) to our purpose.

Definition (Gini–Simpson diversity index). *Let A be a population of agents and* $T = \{t_1, t_2, \ldots, t_n\}$ *the partition that resembles the types in A. Agents expressing the ith type form sets* $t_i \subseteq A$*. Then the Gini–Simpson index is defined as*

$$\text{Gini} - \text{Simpson}(A, T) := 1 - \sum\nolimits_{t_i} \left(\frac{|t_i|}{|A|} \right)^2.$$

The Gini–Simpson index relies on a given partition of individuals into types. In principle, there can be different partitions of a population that may result in different diversity measurements. In groups with epistemic goals, we find it plausible to cluster agents into types based on their beliefs. In our example from table 1, we could sort agents into two types, those accepting the proposition p_3, and those rejecting it, leading to a Gini–Simpson index of $1 - ((2/3)^2 + (1/3)^2) = 4/9$. Later, a clustering algorithm will help us sort agents into types based on their opinions. The diversity values we observe thus depend on the reliability of the chosen clustering algorithm.

For the purpose of this paper, we treat *homogeneity* as the one-complement to diversity, and we will say that a population that is diverse to the degree of d is homogeneous to the degree of $1 - d$.

Diversity measures characterise populations in terms of how frequently a type is expressed. Polarisation measures characterise populations differently, through the distances between individuals. We rely here on the dispersion measure from Bramson et al. (2017), which understands polarisation as the standard deviation of pairwise differences between agents. In comparison to other polarisation measures, dispersion does not require a computationally intensive clustering, while still approximating the values obtained from cluster-based polarisation measures. Like the Gini–Simpson diversity index, cluster-based polarisation measures rely on an antecedent clustering, but would not consider type frequencies but the distances between individuals of different types (Bramson et al. 2017, 122–28). We follow the dispersion definition in Kopecky (2022, sections 4.4–4.14), which is appropriate for the present model.

Definition (Dispersion). *Let* HD *be the Hamming distance between agents' belief systems, or the number of sentences that are evaluated differently. Let A denote the set of agents, n the number of propositions in the decision problem, and* $M(A)$ *the mean distance* *between pairs of agents from the population A:*

$$M(A) := \binom{A}{2}^{-1} \sum\nolimits_{(x,y)\in A, x\neq y} \mathrm{HD}(x,y)/n.$$

Then, with $N = |A|$, *dispersion is defined as the mean absolute standard deviation of pairwise distances:*

$$\mathrm{dispersion}(A) := 2\sqrt{\frac{1}{N}\sum\nolimits_{i\in A}^{N}\left(\frac{1}{N-1}\sum\nolimits_{j\in A, j\neq i}^{N-1}\mathrm{HD}(i,j)/n - M(A)\right)^{2}}.$$

There are interesting implications between agreement, diversity and polarisation, three concepts that characterise the difference of opinion in groups. Dispersion is maximal when the observed population is split into two groups with maximal in-group agreement (pairwise zero-HD) and maximal out-group disagreement (pairwise maximal HD), and it is minimal in the case of complete agreement in the population. Minimal polarisation and minimal diversity thus meet in maximal agreement, but the concepts diverge otherwise. Members of a maximally polarised population will belong to just two clusters and not occupy any middle ground. A fully diverse society is not shaped in such a way. Rather, its agents would scatter into many different types. As previously observed in the literature, rising polarisation implies lowering diversity, or "simplification" (Bednar 2021, 3–4). A population with very diverse opinions cannot have belief polarisation. We will return to these implications in section 3.3.

2.4. Synthetic generation of epistemic group decision problems

How likely are diverse and polarised groups to aggregate inconsistent group opinions through sentence-wise majority voting? We propose an agent-based model to pursue this question. The model consists of two sub-processes. The first sub-process generates a synthetic collection of arguments as the basis of the group's decision problem. The second sub-process samples agents to generate a group with arbitrary degrees of opinion diversity and belief polarisation, as understood by the measures from section 2.3.

Both sub-processes rely on the theory of dialectical structures (Betz 2009). Using this theory, we model decision problems in terms of agreeing on a response toward a set of arguments. The theory describes individual arguments as inferential relations between a set of premises and a conclusion. Arguments described in this way can be dialectically related to other arguments in two ways: defeat and support. One argument defeats another just in case the conclusion of the first is equivalent to the negation of a premise in the second, and one argument supports another just in case its conclusion is equivalent to a premise itself (Betz 2009, 288). A set of arguments together with these two relations make up a *dialectical structure* or, less technically, an *argument map*. An illustration for such a structure is given in figure 1.

The first sub-process of our model synthesises such argument maps under two constraints. First, the argument maps are constructed hierarchically in the sense that some propositions are used as conclusions at the root of the tree while other propositions are only found in more remote leafs. For this hierarchical construction, we designate a subset of propositions as the key propositions of the debate. These propositions can be imagined to be most central to the decision problem. Arguments

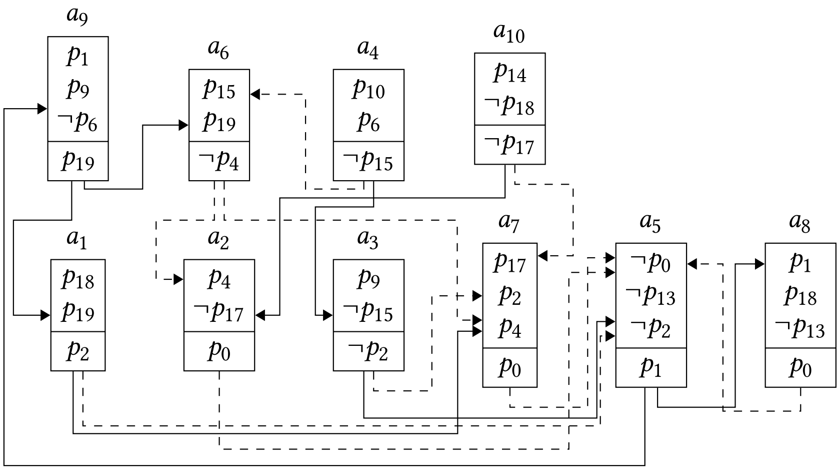

Figure 1. Illustration of a synthetically generated argument map with key statements p_0, p_1 and p_2. Support relations are expressed by solid arrows, defeats by dashed ones.

are generated further away from the roots of the tree by leading to conclusions that are inferentially related to the premises of arguments on lower levels. Second, argument maps are synthesised in such a way that agents have considerable freedom in finding a solution to the decision problem, resulting in situations of epistemic permissibility. In our model, this amounts to there being many beliefs that respect the validity of all presented arguments. Arguments are added iteratively to the map until a specified value of *inferential density* is reached. This parameter can be interpreted as a degree of permissibility and is explained in more detail below.

The example from table 1 in section 2.1 is a minimal instance of our first sub-process and its two constraints. The example contains a single argument and allows multiple equally justifiable but disagreeing responses, three of which were actually maintained in the example. In reality, experts face decision problems with a significantly higher number of propositions and arguments. This is why our model generates complex argument maps as opposed to the minimal example from section 2.1.

Following argument map generation, the second sub-process samples a group of agents with a specified sample size and polarisation or diversity value, depending on the model variant in use. We first describe the composition of individual beliefs, expand on the concept of inferential density, and then introduce our group sampling strategies.

For the description of agents' beliefs, we again rely on the theory of dialectical structures. In the theory, agents' beliefs are expressed by a *belief system*, a mapping from the propositions in all arguments to binary truth values (Betz 2013, 34–6). For example, the beliefs of an agent accepting all premises and the conclusion of a_2 in figure 1 are described by $\{p_4 : \text{True}, p_{17} : \text{False}, p_0 : \text{True}\}$.

In our model, there are two constraints on the beliefs that agents may take in light of an argument map. The first constraint is that every agent assigns a truth value to all propositions from the argument map. This simplification is necessary to allow for voting without abstention, and it implies that we are modelling quite specific scenarios. As all agents are competent to judge all involved propositions, our model is best interpreted as tracking the decision procedure in experts with considerable overlap in expertise. An extension of our model could track decision procedures in multi-disciplinary groups by allowing suspended judgement in the voting procedure. Another extension could model agents having degrees of belief by pursuing probabilistic aggregation (Martini and Sprenger 2017, 185–6).

The second constraint is one of individual consistency. Each agent must hold beliefs that respect the validity of all presented arguments. The number of belief systems that meet this criterion depends on the argument map. Larger argument maps tend to give the agents less freedom in selecting their beliefs as they impose more validity restrictions. In empty argument maps with n propositions under discussion, every agent could have any of 2^n belief systems made up of allocations to True and False. Each argument that is added to the map potentially reduces the number of available belief systems. For example, the argument map consisting of just one argument, $\left(p_1 \wedge p_2\right) \rightarrow p_3$, has $2^3 - 1$ solutions. The one belief system that is unavailable due to violation of the argument's validity is $\left\{p_1 : \text{True}, p_2 : \text{True}, p_3 : \text{False}\right\}$.

As argument maps grow, there is a normalised measure in $[0, 1]$ indicating to what degree agents can choose beliefs freely and to what degree they are predetermined by inferential relations. This measure is called *inferential density*. Argument maps with a density of $D = 0$ impose no constraints on belief choice. At the other extreme, at $D = 1$, only a single validity-respecting belief system remains. The argument map then predetermines all beliefs.

Inferential density is calculated from the number of propositions in the argument map and the number of belief systems that respect the validity of all arguments, but not from the number of agents or arguments. For an argument map τ with n propositions, let the number of validity-respecting beliefs be known as σ_τ. Inferential density is then defined by Betz (2013, 44) as $D(\tau) := \left(n - \log_2 \sigma_\tau\right)/n$. As we show in the appendix, density is the one-complement of the argument map's *normalised information entropy*, or $H_{\mathrm{N}}(\tau)$. We can thus understand inferential density as determining the amount of inferential information encoded in an argument map.

Entropy is useful to further clarify the somewhat loose sense of an agent's "degree of freedom" in selecting its beliefs. Suppose that agents would compose their belief systems by making True/False decisions for each of the propositions in the argument map. Then entropy tells us how many decisions agents make freely, on average, before their remaining choices are predetermined by the argument map. In other words, entropy allows us to estimate how much we can learn about agents' beliefs solely on the basis of the presented arguments. For example, in an argument map with $n = 20$ propositions and a density of $D = 0.4$, we can expect that agents make, on average, $n(1 - D(\tau)) = 20(1 - 0.4) = 12$ True/False decisions before the inferential relations in the argument determine their other beliefs. An argument map with a tighter density of $D = 0.8$ would leave agents with only four such basic decisions on average.

From agents with beliefs that are characterised in this way, our model samples groups with a specified degree of diversity or polarisation. Since we allow multiple agents to have the same beliefs, their belief systems are drawn with replacement from all validity-respecting beliefs. There are usually very many agent samples that can be obtained in this way. For groups of 51 agents as in the experiments presented below, there are often well beyond 10^{200} possible configurations. Expression of diversity and polarisation are not equally distributed within these configurations. Most randomly sampled agent groups would express medium diversity and low polarisation. Our search for groups with specific expressions of diversity and polarisation thus has to be strategic. We describe our group sampling algorithms in more detail in the supplementary materials, but we include a brief summary here. For the diversity variant of the model, we first apply the affinity propagation clustering algorithm (Frey and Dueck 2007) to the collection of all beliefs that respect the validity of the antecedently synthesised map. As we regard membership in these clusters as type expression, we then draw agents from these clusters in such a way that the cluster frequencies result in the desired diversity index. For the polarisation variant of the model, we sample agents following a pyramid scheme of sorts: for a given distance δ, we initially draw a pair of agents with mutual distance δ in their beliefs. We then iteratively draw additional agents of distance δ to a belief system already in the sample until the group contains the desired number of agents. The choice of δ determines the resulting degree of polarisation in the sample.

After synthesising the argument map and sampling an agent group, the model performs a sentence-wise majority vote and verifies whether the individually consistent agents aggregate their beliefs to a consistent group opinion. This process is iterated arbitrarily often in a simulation experiment. At each iteration, the model stores the following information for further statistical analysis: the inferential density expressed by the argument map, either the diversity or polarisation expressed by the sampled agents, and whether the group aggregated a consistent group opinion. In section 3 we present the results from such a simulation experiment.

3. Simulation procedure and results

3.1. Model parameters and main results

In this section we present results from thousands of iterations of both the diversity and polarisation variants of our model. A quantitative analysis of these runs (section 3.2) reveals that the chance of achieving a consistent group opinion through sentence-wise voting drops as opinions diversify. The inconsistency prevalence rises towards medium polarisation but drops for highly polarised groups. In regions of high diversity and medium polarisation, more majoritarian aggregations are inconsistent than consistent. By contrast, regions of low to medium diversity as well as minimal and maximal polarisation show little to no inconsistent aggregation. In our explanation for this initially counter-intuitive pattern (section 3.3) we consider the clustering that groups with different degrees of diversity and polarisation typically exhibit.

A second result is that the inconsistency prevalence is relatively stable across argument maps with different inferential density. Additional information does not by itself bring about consistency in aggregated group beliefs, as long as epistemic permissiveness remains. In fact, highly diverse groups are at a *higher* inconsistency risk as argument maps get more inferentially dense.

These results were gathered from iterations of our model on argument maps with 51 agents and 20 propositions, and their negations, for five points of inferential density (0.4, 0.5, 0.6, 0.7, and 0.8). We use an odd number of agents to simplify the model, as this will not require a decision procedure in the case of a tie. From Betz's formula (2013, 44), we determine the number of validity-respecting beliefs at each inferential density D by solving for x in the equation $D = \left(20 - \log_2(x)\right)/20$. At a density of $D = 0.4$, 20 propositions allow for 4096 validity-respecting belief systems. At a density of 0.6, this number has shrunk to 256, and only 16 validity-respecting belief systems remain at $D = 0.8$. The simulation procedure generates several argument maps per density point and several agent samples for each generated argument map. The supplementary materials contain more details about the exact simulation procedure.

Our data collection ensures that the data is distributed smoothly across the five points of inferential density as well as the full range of opinion diversity and belief polarisation. We collected 10,798 data points in iterations on the diversity variant and 10,722 data points for the polarisation variant of our model. The high number of data points ensures that the results are statistically reliable, even though our model contains random processes in the synthetic generation of argument maps and agent samples.

3.2. Quantitative explorations of many model runs

Figure 2 shows that, in our model, inconsistent majority opinions are much more likely in diverse compared to homogeneous samples. This effect is relatively stable across different degrees of inferential density. Although a majority of diverse groups achieve consistent aggregations at a density of 0.4, the inconsistency risk is still considerable there and in regions of medium density. A rise in inferential density can even increase the prevalence of inconsistent aggregations in diverse groups: at a density of 0.8, a clear majority of highly diverse samples aggregate inconsistent group beliefs. Increasing inferential density does not seem to be a reliable countermeasure to the observed inconsistency risk.

Groups with medium to low opinion diversity rarely aggregate their beliefs to inconsistent group opinions. As we add more inferential information to the synthetically generated argument maps, we start to see inconsistent aggregations in moderately diverse groups more often. Overall, though, inconsistency is a considerable risk only for diverse groups.

Highly diverse groups are at a particular risk of inconsistent aggregation, but groups with high polarisation are not, as figure 3 shows. Highly polarised groups with a dispersion above 0.75 rarely aggregate inconsistent group opinions. This effect is slightly amplified at higher inferential density. Above a dispersion of 0.8, we observe inconsistencies more often at densities of 0.4–0.5 than at 0.6–0.8. The picture differs completely for moderately polarised groups with a dispersion of 0.4–0.6. In almost all areas of inferential density, these groups achieve consistent group opinions relatively rarely. Their share is only noteworthy at a density of 0.8.

Low diversity and low polarisation are both areas of high agreement, which is why we are not at all surprised to see a clear majority of consistent aggregation in these areas. After all, high agreement implies that most agents agree on most issues, and since the agents hold individually consistent beliefs, the aggregated group opinion is highly likely to be consistent as well. We consider this and similar mechanisms in section 3.3.

It is noteworthy that our way of modelling and measuring diversity only allows for few highly diverse samples at a density of 0.8. At this density, only 16 individually consistent belief systems remain as validity-respecting. This naturally limits the number of types expressed in the sample to 16, or usually less. This in turn lowers the

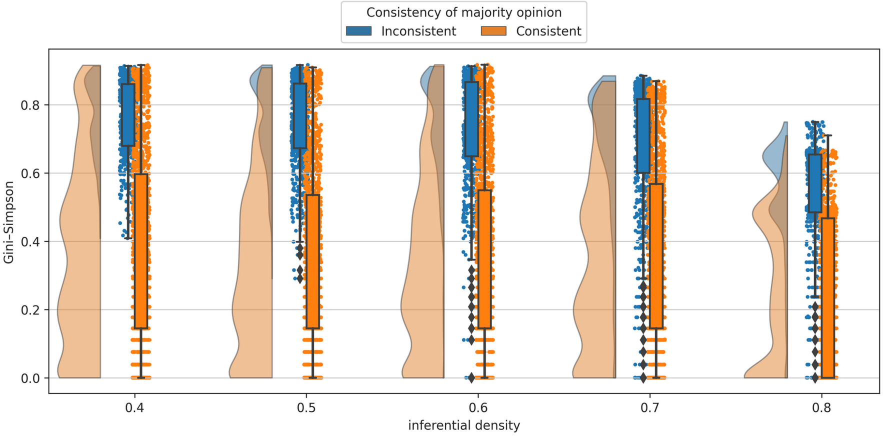

Figure 2. Majority opinion consistency in 10,798 samples of 51 agents with varying diversity, expressed as the Gini–Simpson index, and varying informational influence, expressed as inferential density. Scatter plots show all observations, while the box plots indicate the data points within the 25th to 75th percentiles. As there are about equally many data points in each Gini–Simpson region, a rise in the proportion of consistent observations implies a fall in the proportion of inconsistent ones, and vice versa.

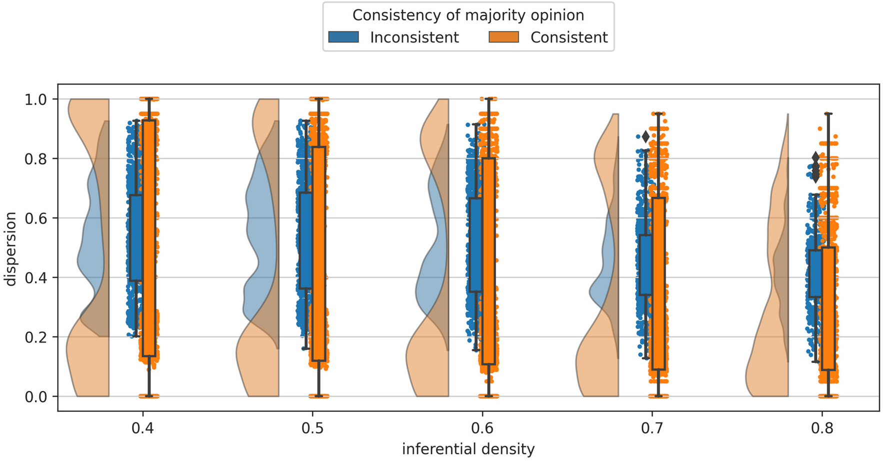

Figure 3. Majority opinion consistency in 10,722 samples of 51 agents with varying polarisation, measured as dispersion, and inferential density. See figure 2 for further description.

maximal Gini–Simpson values we can achieve, as the Gini–Simpson index is sensitive to the number of types. A higher number of types can achieve a higher diversity compared to a lower number of types.

Figures 2 and 3 show a wealth of statistical information about the model, but the results can be expressed more succinctly. First, we offer a summary of the data in table 2. It illustrates that, in our model, the prevalence of inconsistent majoritarian

Table 2. Ratio of consistent group beliefs, across all points of inferential density, depending on diversity and polarisation. As consistency is a binary variable, the ratio of inconsistent aggregations can be derived from the "consistent" column

Gini–Simpson	% consistent
0.00–0.25	95.09
0.25–0.50	82.11
0.50–0.75	44.22
0.75–1.00	28.76
Dispersion	% consistent
0.00–0.25	91.92
0.25–0.50	27.90
0.50–0.75	29.88
0.75–1.00	78.85

aggregation continuously rises as groups diversify. Groups with both very high and very low polarisation are likely to achieve consistent majoritarian aggregation, but moderately polarised groups achieve it in less than a third of all cases.

These two effects can be further quantified using a binary logistic regression analysis. With consistency as the dependent variable and polarisation and diversity as explanatory variables, the logistic models are significant both for the relation between diversity and inconsistency ($\chi^2(1) = 3986, p \ll 0.001, n = 10{,}798$) and for polarisation ($\chi^2(1) = 5125, p \ll 0.001, n = 10{,}722$). The coefficients of these models reveal that the relative probability of achieving consistency drops by 5.8% for every 0.01 gain in diversity (the 95% confidence interval being $[5.6\%, 6.0\%]$). In the polarisation case, the relative probability of achieving consistency rises by 12.8% for every 0.01 change away from 0.5 dispersion to either side ($[12.3\%, 13.2\%]$). As is no surprise in view of figures 2 and 3, Cohen's f^2 indicates a strong effect for diversity (0.40) and an even stronger one for polarisation (0.54).

3.3. Explanations for the success of homogeneous and polarised groups

How can the success of homogeneous and the relatively common failure of diverse groups be explained? There is a seemingly natural and trivial explanation for this effect, but it is not supported by our data. We find a more promising explanation in the degree to which agreeing diverse and polarised agents typically form opinion-based clusters.

The trivial explanation goes: when more than 50% of a population hold exactly the same view, this opinion will be identical to the aggregated majority opinion. Since agents often have identical beliefs in homogeneous and depolarised groups, consistency is brought about trivially in these cases. This trivial factor does not contribute substantially to the observed data. Only 2% of data points in the diversity variant and 15% of observations in the polarisation variant had agent samples in

Table 3. Illustration of a sub-group $a_1, \ldots, a_j, j > m/2$, determining the majority's position on propositions $p_1, \ldots, p_i$ as they share the same view of these propositions (marked by "$+$"). The other judgements are left open (indicated by "?")

Opinion of	p_1	p_2	$\cdots$	p_i	p_{i+1}	$\cdots$	p_n
a_1	$+$	$+$	$\cdots$	$+$	?	$\cdots$	?
a_2	$+$	$+$	$\cdots$	$+$	?	$\cdots$	?
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
a_j	$+$	$+$	$\cdots$	$+$	?	$\cdots$	?
a_{j+1}	?	?	$\cdots$	?	?	$\cdots$	?
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
a_m	?	?	$\cdots$	?	?	$\cdots$	?
Majority	$+$	$+$	$\cdots$	$+$	?	$\cdots$	?

which an absolute majority shared the exact same beliefs. The relatively high share in the polarisation variant can be explained by the fact that groups with very high dispersion can only be sampled as two groups holding exactly opposing views. This is true for the dispersion measure but would likewise hold for other polarisation measures such as *group divergence* (Bramson et al. 2017, 125). Since we always sampled an odd number of agents, one of these two groups is home to more than half of the agents. When we factor out these cases with maximal polarisation, the trivial explanation accounts for only 10% of our data.

We find a more promising explanation in the low number of opinion clusters that both polarised and highly agreeing groups exhibit. Highly agreeing populations form a single opinion-based cluster, and bipolarised populations, by definition, form two clusters. This clustering dissipates as groups diversify, leading to more clusters that have fewer members (as displayed in figure 4).

The agents in large clusters can determine the majority's view on a subset of issues. Even if the opinions in such a cluster do not completely overlap, they will usually agree on a considerable number of issues – or they could not form a cluster. And as each agent holds a consistent opinion, the (partial) belief system formed by the cluster's agreement will also be consistent. In highly polarised and homogeneous groups, there is a high chance that this mechanism will indeed fix the majority view on at least part of the propositions (see table 3 for an illustration of this mechanism). On the other hand, a diverse group is far less likely to profit from this mechanism.

In the presence of large opinion clusters, a potential inconsistency would have to be introduced through one of the sentences that the cluster does not agree on. But their introduction is far from guaranteed, especially in environments with low validity constraints: these uncertain epistemic situations allow many group opinions to be consistent. When highly agreeing clusters determine all but a few judgements of the group as a whole, given uncertainty, many extensions of the partially settled majority opinion will be consistent as well, by mere statistical likelihood. This likelihood of achieving consistency by chance drops as fewer opinions remain consistent at higher inferential density.

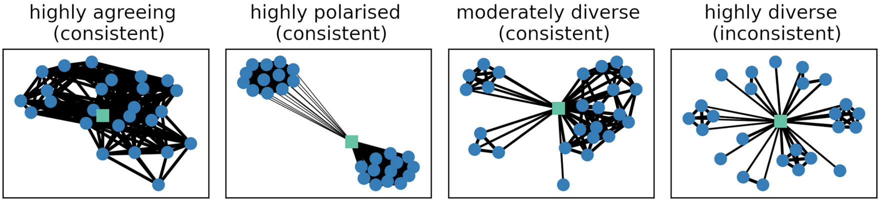
highly agreeing (consistent)
highly polarised (consistent)
moderately diverse (consistent)
highly diverse (inconsistent)

Figure 4. Four samples with 25 agents each. The majority opinions are printed as green squares and the agents as blue circles. Relative node position is a rough indicator of distance. All edges between agents and the majority opinion are plotted and weighted by distance, but edges between agents are only plotted if they disagree about 0, 1, or 2 of the 20 propositions. From left to right, the agents group into an increasing number of clusters. The highly agreeing group consists of only one cluster and the bipolarised sample has two clusters. As opinion diversity rises in the group, more and more clusters become discernible until none can be made out. (Color online.)

This consideration also explains how a rise in inferential density increases the inconsistency risk for moderately and highly diverse groups. While diverse groups with little opinion overlap might find one of the many consistent opinions that low-density environments allow, this strategy, guided more by chance than systematicity, will become less accurate as the number of consistent opinions drops in more dense argument maps.

4. Implications of inconsistent group opinions for expert advice

Following the advice of homogeneous expert groups can negatively affect the legitimacy of subsequent policy making if that homogeneity is not an adequate reflection of the available evidence – and a call to opinion diversity in expert groups is a natural response to this reasonable fear. But inconsistency is another potential source of legitimacy flaws, and expert groups with high opinion diversity are particularly affected by it when pursuing majority voting as part of their aggregation procedure. This is an under-appreciated risk in uncertain information environments, or when the evidence is permissive. We found it to be particularly intricate as it manifests itself despite individual consistency and could not be eliminated in environments with a higher availability of inferential information.

We now consider the implications of these results for the paradigmatic scenarios of expert group decision-making from section 1.

As inconsistencies might escape an external party, such as in the *expert poll* scenario outlined above, we cannot recommend pursuing majoritarian aggregation of diverse opinions under condition of uncertainty without proper reflection of the outcome.

Our results do not show that diverse expert groups would necessarily issue inconsistent advice in the real world, particularly if they become aware of them. What the results indicate is that, when faced with diverse opinions and permissive evidence, setting up reliable aggregation procedures becomes a significant issue for expert groups, such as in the *expert meeting* scenario. After all, they are less likely to be able to rely on a majority vote. An alternative aggregation procedure is described by the Lehrer–Wagner model (Lehrer and Wagner 1981). Under favourable conditions and upon sufficient iteration, it is guaranteed to achieve unanimity and thereby avoid inconsistencies. However, this model is considerably more demanding than majoritarian aggregation. In particular, it would require experts to assign precise weights to the judgements of all other involved peers, and usually requires several iterations to arrive at group consensus. In general, real-world groups such as in the *expert meeting* scenario are more likely to require additional time to reflect on aggregation procedures the more diverse they become. This can affect the difficulty of their epistemic group problem solving as a whole. But these groups do have interesting options available to them, even if they do not use an aggregation method with formal guarantees such as the Lehrer–Wagner model. These options include issuing separate sets of recommendations that each reflect a portion of the diverse group, or they could limit their recommendation to those parts of the issue on which they find a consistent majoritarian opinion.

Unfortunately for the *expert deliberation* scenario, we were unable to find evidence that the mere accumulation of inferential information reduced the inconsistency risk for diverse groups. In fact, as long as the evidence remained permissive, rising inferential density *increased* the chance of inconsistencies in the upper diversity regions.

This does not imply that deliberation is entirely futile. We only observed the voting result after expert deliberation had presumably taken place, but we did not investigate deliberative processes themselves. Although only in a minority of cases, we *did* observe consistent majoritarian aggregation in diverse groups facing decision problems of high inferential density. This raises a worthwhile problem for future research: are there specific deliberative behaviours that help expert groups achieve consistency as inferential density rises? And are there other behaviours that are detrimental to that goal? At the moment, we find the hypothesis that some deliberative behaviours could be particularly conducive to consistency plausible in light of previous research that found substantially different agreement and polarisation dynamics for different types of deliberative behaviour (Betz 2013; Kopecky 2022).

On a related note, it is important to emphasise that the decision problems in our model are static and mutually independent. The model does not track changes to beliefs that agents choose, the arguments underlying the decision problem, or other *dynamic* aspects of belief aggregation. Such dynamic aspects are studied in the literature (e.g., Dietrich 2021) and it seems worthwhile to pursue these aspects further. For example, one could look for optimal strategies to retain the consistency of group opinions if new evidence is introduced or the group composition changes.

We did not find a penalty to consistent aggregation in polarised groups, and we do not see a reason to avoid experts with high belief polarisation – if the polarisation is a consequence of experts following diverging yet consistent paths a permissive set of evidence provides. However, belief polarisation is only one of several ways in which agents can move apart, and our model did not include other types of polarisation that disrupt deliberation and decision-making, such as affective polarisation or ideological alignment (e.g., Iyengar and Westwood 2015).

The high prevalence of inconsistent majoritarian aggregation in very diverse groups could be seen as a trade-off between maximising diversity (and thereby risking inconsistency) on the one hand and minimising risk of inconsistencies (and thereby sacrificing diversity) on the other. But this is a trade-off only in theory. In practice, the composition of expert groups is not determined through the public's or policy makers' desire for diversity and consistency, but rather through academic and epistemic factors. We see the value of our results not in motivating the engineering of expert group composition, but rather in understanding the consequences to expert advice that given compositions have.

As the public faces hard questions it is an understandable desire to obtain consistent and well-informed recommendations that reflect all the diverse opinions consistent with the evidence. In situations that involve permissive evidence and considerable time constraints, this desire may not always be satisfiable. Instead, citizens and policy makers should be aware that experts might offer conflicting or incomplete advice when such exceptional conditions hold, and make provisions for decision-making under uncertainty if the issue cannot be immediately resolved through expert advice. Our data indicates that relaying decision-making to expert opinion might have limits where decisions must be made without delay but the evidence permits diverse and equally justifiable recommendations. A failure to recognise these challenges might put experts in the difficult situation of being expected to solve impossible epistemic decision problems while simultaneously being blamed for not actually solving them.

5. Conclusion

Are groups with highly diverse beliefs better epistemic problem solvers, and polarised groups always worse? Not necessarily – specifically, when pursuing majoritarian aggregation under uncertainty and permissive evidence, diverse groups yield inconsistent outcomes more often than homogeneous and polarised groups. Decision-making can be more difficult for diverse groups in these scenarios, not least because evidence accumulation does not necessarily improve their situation.

There are difficult but worthwhile questions related to the risk of inconsistent aggregation. Will we be able to identify consistency-conducive types of deliberative behaviour? If consistency cannot be achieved, should experts issue separate, individually consistent, minority recommendations? Or should they explicitly restrict their recommendations to issues backed by a consistent majority? And how should policy makers and the public react to the described difficulties? Should expert advice be superseded in the case of inconsistent or inconclusive recommendations, such as by overarching agreement on cultural or moral ideals?

Basing public decision-making on homogeneous expert groups incurs a legitimacy risk if the evidence would allow for diverse opinions. Increasing opinion diversity is a legitimate request in these situations, but comes with its own set of problems. The difficulties faced by diverse groups should not be taken as evidence for poor performance but should rather be taken to indicate just how complex it can be to find consistent advice on time-critical issues when the evidence permits multiple justified approaches.

Acknowledgements. We are very grateful to have received constructive and most useful comments from two anonymous peer reviewers and the editor in chief of this journal.

An earlier version of this paper was presented at the Institutional Epistemology Workshop (INES 2023) at the University of Helsinki and at a workshop on reasoning with imperfect information in social settings at Scuola Normale Superiore in Pisa. We thank the audiences for discussing our work, and Raul Hakli, Säde Hormio, Maria Lasonen-Aarnio, Samuli Reijula and Kristina Rolin (Helsinki) as well as Mario Piazza, Caterina Sisti, Matteo Tesi and Pietro Vigiani (Pisa) for organising the workshops.

We also gratefully acknowledge support by the state of Baden-Württemberg through bwHPC, the state's high-performance computing consortium.

References

Bednar, Jenna. 2021. "Polarization, Diversity, and Democratic Robustness". *Proceedings of the National Academy of Science* 118 (50):e2113843118. DOI: https://doi.org/10.1073/pnas.2113843118.

Betz, Gregor. 2009. "Evaluating Dialectical Structures". *Journal of Philosophical Logic* 38:283–312. DOI: https://doi.org/10/cxrbhh.

Betz, Gregor. 2013. *Debate Dynamics: How Controversy Improves Our Beliefs*. Berlin: Springer. DOI: https://doi.org/10/d3cx.

Bramson, Aaron, Patrick Grim, Daniel J. Singer, William J. Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman. 2017. "Understanding Polarization: Meanings, Measures, and Model Evaluation". *Philosophy of Science* 84 (1):115–159. DOI: https://doi.org/10.1086/688938.

Dietrich, Franz. 2021. "Fully Bayesian Aggregation". *Journal of Economic Theory* 194:105255. DOI: https://doi.org/10.1016/j.jet.2021.105255.

Frey, Brendan J. and Delbert Dueck. 2007. "Clustering by Passing Messages between Data Points". *Science* 315 (5814):972–6. DOI: https://doi.org/10.1126/science.1136800.

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. DOI: https://doi.org/10.1086/701070.

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 Science* 101 (46):16385–9. DOI: https://doi.org/10.1073/pnas.0403723101.

Iyengar, Shanto and Sean J. Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization". *American Journal of Political Science* 59 (3):690–707. DOI: https://doi.org/10.1111/ajps.12152.

Kopecky, Felix. 2022. "Arguments as Drivers of Issue Polarisation in Debates among Artificial Agents". *Journal of Artificial Societies and Social Simulation* 25 (1):4. DOI: https://doi.org/10.18564/jasss.4767.

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

List, Christian. 2005. "The Probability of Inconsistencies in Complex Collective Decisions". *Social Choice and Welfare* 24:3–32. DOI: https://doi.org/10.1007/s00355-003-0253-7.

List, Christian and Philip Pettit. 2002. "Aggregating Sets of Judgments: An Impossibility Result". *Economics and Philosophy* 18 (1):89–110. DOI: https://doi.org/10.1017/S0266267102001098.

Martini, Carlo and Jan Sprenger. 2017. "Opinion Aggregation and Individual Expertise". In *Scientific Collaboration and Collective Knowledge: New Essays*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg. Oxford: Oxford University Press.

Page, Scott E. 2011. *Diversity and Complexity*. Princeton, NJ: Princeton University Press. DOI: https://doi.org/10.1515/9781400835140.

Pettit, Philip. 2001. "Deliberative Democracy and the Discursive Dilemma". *Philosophical Issues* 11 : 267–99.

Pöyhönen, Samuli. 2017. "Value of Cognitive Diversity in Science". *Synthese* 194:4519–40. DOI: https://doi.org/10.1007/s11229-016-1147-4.

Richardson, Henry S. 2002. *Democratic Autonomy: Public Reasoning about the Ends of Policy*. Oxford: Oxford University Press.

Shi, Feng, Misha Teplitskiy, Eamon Duede, and James A. Evans. 2019. "The Wisdom of Polarized Crowds". *Nature Human Behaviour* 3:329–36. DOI: https://doi.org/10/c286.

Tuomisto, Hanna. 2010. "A Consistent Terminology for Quantifying Species Diversity? Yes, It Does Exist". *Oecologia* 164:853–60. DOI: https://doi.org/10.1007/s00442-010-1812-0.

Vallier, Kevin. 2022. "Public Justification". In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Stanford, CA: Stanford University Press. https://plato.stanford.edu/archives/win2022/entries/justification-public/.

Zollman, Kevin J. S. 2010. "The Epistemic Benefit of Transient Diversity". *Erkenntnis* 72:17–35. DOI: https://doi.org/10.1007/s10670-009-9194-6.

Appendix

Supplementary materials

A repository at https://zenodo.org/record/10580623 contains Jupyter notebooks to run the experiments described in this paper and analyse the data.

The relation between inferential density and normalised information entropy

In section 2.4 we claim that inferential density is the one-complement to normalised entropy, but we have delayed the justification for this claim to this appendix.

For the calculation of normalised entropy, let $p(i)$ denote the probability that an agent would randomly pick a belief system i out of 2^n possible belief systems in light of an argument map with n propositions.

We know that some, but not all, of the 2^n belief systems will respect the validity of all the arguments presented in the argument map τ. Let Γ_τ denote this set of belief systems and σ_τ its size, $\sigma_\tau = |\Gamma_\tau|$. Since the agents in our model will only accept validity-respecting beliefs, we can further characterise $p(i)$:

$$p(i) = \begin{cases} 0 & \text{if } i \notin \Gamma_\tau, \\ 1/\sigma_\tau & \text{if } i \in \Gamma_\tau. \end{cases}$$

We can use this knowledge to transform the normalised entropy $H_\mathrm{N}(p)$ for our $p(i)$. With $N = 2^n$, we observe:

$$H_\mathrm{N}(p) = -\sum\nolimits_i \frac{p(i)\log_b p(i)}{\log_b N} = -\underbrace{\left(\frac{(1/\sigma_\tau)\log_2(1/\sigma_\tau)}{\log_2 2^n} + \cdots + \frac{(1/\sigma_\tau)\log_2(1/\sigma_\tau)}{\log_2 2^n}\right)}_{\text{repeated } \sigma_\tau \text{ times}}$$

Since $p(i) = 0$ for all non-validity-respecting belief systems, these drop out of the sum. The sum over the remaining validity-respecting beliefs can also be written as a product:

$$H_\mathrm{N}(p) = -\frac{\sigma_\tau(1/\sigma_\tau)\log_2(1/\sigma_\tau)}{\log_2 2^n}.$$

Basic properties of the binary logarithm then allow this transformation:

$$H_\mathrm{N}(p) = -\frac{\sigma_\tau(1/\sigma_\tau)\log_2(1/\sigma_\tau)}{\log_2 2^n} = -\frac{\log_2(1/\sigma_\tau)}{n}$$

$$= -\frac{-\log_2\sigma_\tau}{n} = \frac{\log_2\sigma_\tau}{n} = 1 - D(\tau).$$

We thus say that the inferential density of an argument map τ is the one-complement to its normalised entropy, or $D(\tau) = 1 - H_\mathrm{N}(\tau)$.

Cite this article: Kopecky, Felix and Gregor Betz. 2025. "Inconsistent Belief Aggregation in Diverse and Polarised Groups." *Philosophy of Science* 92 (1):40–58. https://doi.org/10.1017/psa.2024.29