Philosophical Studies
https://doi.org/10.1007/s11098-023-02059-6

Argumentation-induced rational issue polarisation

Felix Kopecky[1]

Accepted: 7 October 2023

Abstract
Computational models have shown how polarisation can rise among deliberating agents as they approximate epistemic rationality. This paper provides further support for the thesis that polarisation can rise under condition of epistemic rationality, but it does not depend on limitations that extant models rely on, such as memory restrictions or biased evaluation of other agents' testimony. Instead, deliberation is modelled through agents' purposeful introduction of arguments and their rational reactions to introductions of others. This process induces polarisation dynamics on its own. A second result is that the effect size of polarisation dynamics correlates with particular types of argumentative behaviour. Polarisation effects can be soothed when agents take into account the opinions of others as premises, and they are amplified as agents fortify their own beliefs. These results underpin the relevance of argumentation as a factor in social-epistemic processes and indicate that rising issue polarisation is not a reliable indicator of epistemic shortcomings.

Keywords Social epistemology · Polarisation · Argumentation · Deliberation · Agent-based models (ABM) · Opinion dynamics · Epistemic rationality

1 Introduction

Many explanations for the rise of polarisation among humans are compatible with, or even suggest, the involvement of epistemically irrational behaviour. Candidates include preventing exposure to the views of others (Mutz, 2002), a confirmation bias toward one's own views and a disconfirmation bias toward contrary positions (Taber & Lodge, 2006), or (ideologically) motivated reasoning (Kahan, 2013). These explanations support irrationality as contributing to polarisation given that *rationality* can be understood not just as having a coherent set of mind, but also in terms of responsiveness to evidence (see Fogal & Worsnip, 2021, for a recent discussion of this idea). Purposefully ignoring evidence to the contrary of one's view or biasing

✉ Felix Kopecky
f.kopecky@kit.edu

[1] Karlsruhe Institute of Technology, DebateLab, Karlsruhe, Germany

Published online: 02 December 2023

evidence evaluation inhibits correctly responding to the evidence, and therefore can be seen as irrational.

But is polarisation always avoidable when agents meet the conditions of epistemic rationality? In other words, is rising polarisation among deliberating agents necessarily evidence of irrationality? Singer et al. (2019) say “No”. In simulations on their agent-based debate model, polarisation rises even when all agents comply with the rationality criteria required by their model. While Singer et al. would not deny that irrationality can be a contributor to rising polarisation, their data suggest that irrationality is not a necessary condition for rising polarisation. Deliberating agents can do the best they can – and still end up polarised. Section 2 of the present paper is about what it means for agents to polarise and the diverse ways in which they can do so.

The severe limitation to agent memory in Singer et al.’s model raise concerns about how convincing the case for rational polarisation actually is. Are agents that can remember only a handful of propositions really engaging in rational debate? I discuss worries about limiting artificial agents in their ability to interact with their epistemic surroundings in Sect. 3.

The agent-based debate model presented in Sect. 4 remedies this situation. It is a model based on Betz’s (2009; 2013) theory of dialectical structures. In the model, agents with perfect memory purposefully exchange arguments, understood as premise-conclusion structures, and respond rationally to arguments presented by others. Simulations on this model support the case for epistemically rational issue polarisation. As a matter of principle, epistemic rationality constraints do not prevent deliberating agents to polarise in the specific sense of issue polarisation (Sect. 5). This possibility raises new questions about what, if anything, is wrong with issue polarisation, and which interventions its occurrence requires (Sect. 6).

The results go beyond that. Simulations on this new model reveal how polarisation dynamics differ substantially depending on which argumentative behaviour the agents pursue. Polarisation effects are soothed as agents construct arguments from premises shared with others, and are amplified through arguments that unilaterally strengthen one’s own position. Besides the striking impact of reasoning with shared beliefs, the results point us to the non-obvious but profound social impact of continuous unilateral belief fortification. As I discuss in Sect. 6, these results underpin the relevance of argumentation to matters in social epistemology.

2 Polarisation concepts and their relevance for model design and evaluation

Recent contributions to sociology suggest that *polarisation* does not describe a single phenomenon, but that it is best understood as a cover term for a collection of concepts (Mason, 2013, 2015; Iyengar et al., 2012). There are three specific subtypes that describe different ways in which agents can polarise: affective polarisation, polarisation of issue positions and group polarisation.

Affective polarisation is characterised by increasing animosity between groups each defined by a shared identity.[1] For example, recent polling data suggests that Republicans and Democrats in the United States become increasingly unlikely to socially interact, such as in marriage or friendship (Pew Research Center, 2014, 2017) – a clear indicator of affective polarisation in the United States. In the global context, some countries see affective polarisation dynamics comparable to the US, but others have recently experienced stagnation or a fall in polarisation (Boxell et al., 2022). Boxell et al. (2022, §2) understand affective polarisation as the aggregated differences in respondents' affect toward their preferred political party compared to other parties. Respondents with a high difference in affect would contribute to higher polarisation, while respondents with no difference would contribute to depolarisation. Other measures of affective polarisation include implicit association or behavioural tests (Iyengar et al., 2019, pp. 131–133). All of these measures are based on respondents' sympathy toward other individuals – but they do not poll their issue positions.

Affective polarisation does not necessarily imply divergence on specific issues (Mason, 2015, p. 128). The second polarisation concept, *issue polarisation*, determines how much on-topic beliefs move apart concerning a specific issue. Bramson et al. (2017) collect different interpretations of this phenomenon: it could mean a rise of variance among held beliefs in a population, but maybe the most comprehensive interpretation is how clusters form and grow apart (Bramson et al., 2017, §2.5–2.9). Issue polarisation understood in this way is best characterised by the belief-based formation of groups that become more cohesive internally while simultaneously diverging from other, likewise increasingly cohesive, groups.

Issue polarisation is about agents' stances toward on-topic claims, while affective polarisation concerns agents' attitudes toward other agents. Recognising affective polarisation as a distinct phenomenon elucidates that controversy and division in humans can not always be comprehensively described with reference to their difference *of opinion*. There are cases in which their difference *in sympathy* is essential.

The third polarisation concept, *group polarisation*, runs orthogonal to the distinction of the two previous kinds. It captures the effect that groups move to more extreme positions than its member initially had (see, e.g., Myers, 1975; Sunstein, 2002). This phenomenon could be understood both in the issue or affective sense of polarisation, but is not investigated here.

Distinguishing affective and issue polarisation is relevant for the design and evaluation of computational models, since all polarisation models involve a choice which of these to implement. This choice determines the real-world events that we can hope to better understand through modelling. The majority of models in the philosophical literature, and all models discussed in Sect. 3, track agents' issue positions.

What are real-world examples for the occurrence of issue polarisation? The public can polarise over political issues, even along party lines. But the concept applies

[1] The terminology used in the literature is diverse. "Social" is the term used by Mason, while Iyengar et al. use the term "affective polarisation". I treat these terms as synonymous.

not only to political cases. The history of science has ample examples of scientists converging with in-group members but diverging from the opinions of other groups, such as in geology (Hallam, 1989) or in Lyme disease research (O'Connor & Weatherall, 2018, §2). Disagreements between judges in a judicial panel can polarise during their epistemic quest to establish the guilt of a defendant or the constitutionality of a law. Philosophical debates can polarise in the issue sense as well: in fact, we regularly give names to members of groups which, to varying degree, converge internally but diverge externally ("externalists" and "internalists", "empiricists" and "rationalists", "moral realists" and "constructivists", etc.).

3 Limitations in polarisation models

Can opinions on issue positions polarise in deliberating artificial agents? Can they do so while complying with epistemic rationality demands, such as belief coherence and responsiveness to evidence? Singer et al. (2019) answer these questions affirmatively. In their computational model, agents are equipped with a belief system that stores *n* reasons out of all reasons available in the simulated world. Reasons are represented by real numbers indicating their strength and sign. If they have positive sign, they lead agents to belief the single proposition under discussion. Reasons with a negative sign lead them to disbelief it. Whether an agent beliefs the proposition and the strength of its conviction are determined by the sum of all reasons it currently possesses. Agents constantly receive new reasons from the world or through unbiased, public communication with other agents. As their memory is limited, a previously possessed reason must be dropped for every one that enters their memory. Singer et al. specifically claim epistemic rationality for their "coherence-minded" strategy of forgetting, in which agents forget the reason that least supports their current opinion.

It is due to these memory limits that their case for epistemically rational issue polarisation is not entirely convincing. Concerns arise when considering how limited the agents are in epistemically interacting with the world, particularly considering the limited memory of seven reasons in the main experiment, out of 500 reasons available in total. The polarisation effect weakens as agents have larger memory and vanishes under condition of perfect memory (Singer et al., 2019, Figure 1, p. 2250). There are at least three reasons why models with severe memory restrictions do not provide straightforward support for the possibility of rational polarisation:

1. If agent memory is limited to a very low number, such as 7, then it is hard to imagine how a meaningful discussion could take place among the agents. Just try to imagine academics discussing a talk at a conference under such limits, or what consequences such a limited memory would have for everyday doctor-patient interactions. This is not to say that agents with limited memory cannot fulfil *some* criteria for epistemic rationality, such as being free of conflicting beliefs, or basing their views solely on their evidence. But these necessary conditions are trivially met even by belief systems that can not participate in debate in a meaningful way, such as the minimally coherent opinion ∅. Put more generally,

many processes to form rational beliefs require access to a substantial amount of memory items, and agents with sparse memory are unable to activate these processes, thus being unable to attain rational beliefs.

2. Singer et al. motivate these memory limits by referring to the psychological literature, which suggests that humans can retain four or seven items in memory – this limit is known as "magic number four" (or "seven") in psychology. It is vital to note that this limit covers items in *short term memory* (STM) only and *does not cover other types of memory*. Typical STM tasks include memorising a phone number or e-mail address for less than half a minute, adding two numbers or reading and comprehending a single sentence (Jonides et al., 2008). Clearly, deliberation requires input from other types of memory, such as important facts that agents learned in vocational training or graduate school and will retain in their memory for the rest of their lives.
3. But even if limitation of agent memory was permissible, the model does not explain why agents can not draw on other deliberative and evidential resources, such as notebooks. In *Twelve Angry Men*,[2] there are several instances of jurors referring to their notes from the court hearing. And as deliberations are increasingly conducted on digital platforms, agents can refer to even more such resources. In fact, when rational agents realise their memory to be too limited to accommodate all pertinent evidence, they would react by referring to memory enhancing or externalisation techniques, such as taking notes, or by collectively charting the debate on a white board. Modelling agents not to have access to notes, text books, or other resources external to their memory does not adequately model the epistemic abilities of rational agents.

Many models of polarisation dynamics in the literature depend on limiting agents' access to their epistemic surroundings. In O'Connor and Weatherall (2018), agents hold a belief in [0, 1] and update it by conditionalising on the beliefs of other agents depending on whether they trust them (O'Connor & Weatherall, 2018, pp. 861–864). Theirs is not a model under epistemic rationality (2018, p. 857), but their polarisation effect also vanishes when agents are not restricted in their epistemic interaction with the world. Their limiting factor is mistrust, which makes agents unresponsive to signals broadcast by others they do not trust (2018, pp. 866–868).[3]

[2] *Twelve Angry Men*, a 1955 play by Reginald Rose, follows a jury debating a murder case in a US court. Singer et al. (2019) treat this debate setting as a prototypical case. On first appearance, a lot of circumstantial evidence seems to support the defendant's guilt, but through continued debate and by going through different stages of agreement, the jurors come to the conclusion that the evidence is not decisive after all and they find the defendant not guilty due to reasonable doubt, meaning they end up non-polarised. The play is re-assuring to an optimistic outlook on the power of argumentation, because arguments, rather than manipulation, aggression or social ties are portrayed as the tool with which the jurors arrive at their conclusion.

[3] There is an interesting difference between these two models: Singer et al.'s agents base their opinion on information and only indirectly on what others believe through the reasons they communicate to them. Not all computational polarisation models expose their agents to "informational" influence. O'Connor and Weatherall (2018) have their agents adapt their own beliefs directly in light of other agents' beliefs, particularly those closely related to them. This kind of influence is known as "social" influence (see Proietti and Chiarella (2021, p. 1–2) or Burnstein and Vinokur (1977) for a discussion of this distinction).

Singer et al.'s model is related to a sociological model by Mäs and Flache (2013). A comparison between these two highlights the different applications of computational modelling in sociology and epistemology. This points us to the specific and substantial contribution that philosophical models can offer. Like Singer et al., Mäs and Flache determine their agents' beliefs through aggregating a set of currently possessed reasons, although they normalise beliefs to [0, 1] and interpret these reasons to be "arguments" (Mäs & Flache, 2013, §1.3.2). Just like in Singer et al., agents can remember a limited number of reasons (6 in Mäs & Flache, 7 in Singer et al.). But there are noteworthy differences. Mäs & Flache do not allow public, unbiased communication among their agents. Instead, the chance of communication occurring between agents depends on whether they hold similar beliefs, and communication involves the exchange of reasons only for the agents partnered at this stage. Agents also manage their memory differently in Mäs and Flache (2013). They forget the reason they held for the longest time, irrespective of how well it supports their current opinion. Although the model shares many formal similarities with Singer et al.'s, it could not be used to make the case for epistemically rational polarisation. Public and unbiased communication as well as coherence-minded updating are features in Singer et al.'s model that allow them to approach the question of whether polarisation is possible under condition of epistemic rationality. The research question is a rather different one for Mäs and Flache (2013): they are looking for explanations of bi-polarisation, irrespective of whether it is brought about rationally. In a bi-polarised population, positions on both ends of a spectrum are each upheld by about half of the population, but few if any take the middle ground. A bi-polarised outcome is by no means necessary to show that a significant rise in polarisation is possible under epistemic rationality.

Asking whether polarisation can occur among agents that exhibit epistemically rational behaviour is a clearly delimited research question with substantial philosophical interest. The model presented in the remainder of this paper pursues this interest – but it does not rely on limitations to how agents evaluate their epistemic surroundings.

4 Modelling debates through exchanges of arguments

The present model is based on the theory of dialectical structures (TDS, Betz, 2009) and is related to an earlier computational model built on this theory (Betz, 2013). The present model is presented in full in this section, including the parts where it deviates from earlier implementations: the possibility for agents to withhold judgement, extension of the sentence pool, tree-like debate growth and, of course, the mechanisms and measures to track polarisation in TDS models. But first, Sect. 4.1 addresses the question what makes argumentative models interesting for computational approaches in social epistemology.

4.1 Argumentation as a social-epistemic phenomenon

Agents in this model interact through argumentation: they construct valid arguments, introduce them to a public debate forum and react to arguments others have introduced in a way that ensures their beliefs remain rational. But why should philosophers and social epistemologists be interested in argumentation? Besides the fact that it provides a model of polarisation dynamics under condition of epistemic rationality without several of the common limitations in extant models, is there a substantial reason to be interested in argumentation other than the simulation results of this model?

Following a popular interpretation of how argumentation fits into (social) epistemic processes (Dutilh Novaes, 2021, §1, §4.5), arguments are transceivers between belief systems. Their purpose is to make others aware of how an agent reasoned and to which conclusion it arrived. This use of argumentation is abundant in deliberative processes, such as in court, academic deliberation, or in parliament. In all of these deliberative institutions, agents rely on arguments to engage with the views of others and explain their own. Argumentation models are conducive to understanding polarisation dynamics in these deliberative contexts.

Some go beyond argumentation's role in multi-agent deliberation and add that argumentation has fundamental functions in our individual epistemic lives. Mercier and Sperber (2011) present an account in which reasoning is an inherently argumentative process. They think that argumentation is a fundamental activity in humans with universal applications considering our rational activities. From their point of view, the question of how arguments fit into socially epistemic practices is a trivial one: reasoning just is producing arguments. Consequently, argumentation could not only model deliberation adequately, but reasoning processes in general.

Cartwright's (2013) theory of evidence is another case in which argumentation occupies a fundamental epistemic role. In her theory, the existence of an argument determines whether something is evidence for a hypothesis. For that to be the case, a suitable proposition about that piece of evidence must be part of an argument. In Cartwright's words: "*e* is evidence for hypothesis *h* relative to a good argument A [...] if and only if *e* is a premise in A, which is itself a good argument for *h*" (Cartwright, 2013, p. 5).

Cartwright's theory of evidence aligns well with social-epistemic practices related to evidence sharing. Evidence, after all, is rarely shared in isolation, but to support a claim through maintaining an inferential relation from a statement about the evidence to a claim. Putting forward pieces of evidence $e_1, e_2, ..., e_n$ to support p is straightforwardly represented by an argument with premises about $e_1, e_2, ..., e_n$ and the conclusion p.[4] And Cartwright's theory also captures disagreements about

[4] Humans, of course, usually deviate from providing arguments in formulaic language, like $(p_1 \wedge ... \wedge p_n) \implies c$. But the premise-conclusion structure of reasoning about evidence can also be found in real-world contexts, such as when the conclusion is stated by one, but the premises by a second participant. Arguments are not necessarily present verbatim in all instances of sharing evidence, but they serve as an adequate *abstract representation* of this process.

evidence in terms of argumentative behaviour: for example, an agent does not need to reject the truth of a premise, such as "the defendant possessed a knife at the time of the murder", but it can reject the inference from its truth toward the guilt of the defendant (e.g., because such a knife is widely available).

Both Mercier and Sperber (2011) and Cartwright (2013) advance substantial theories about what reasoning and evidence fundamentally are, but nothing in the following requires accepting strong readings of these theories. The benefits of studying argumentation in social epistemology can be recognised without buying into these commitments. Instead, we can consider argumentation as a useful *representation* of deliberation, reasoning and evidence exchange, while leaving the ontological questions open.

4.2 The debate forum

The model uses a continuous debate forum, in which the agents are aware of all introduced arguments and the belief systems of all other agents. This is similar to the setting in *Twelve Angry Men*, in which all 12 jurors can hear and respond to all arguments in the debate. Debates in the model start with a pool of n atomic propositions, which is extended to m atomic propositions through introduction of new sentences during the debate. For each atomic proposition p, both p and $\neg p$ are available as premises or as the conclusion of an argument. Consequently, agents can draw on $2n$ premises initially and $2m$ premises when all sentences have been introduced. A subset of atomic propositions is deemed to contain *key statements* of a debate. Their role is explained in the section on argument introduction (Sect. 4.4.1).

Changes to the debate forum over time are tracked in terms of *debate stages*, referred to under the variable τ_i for stage i. A debate stage consists of the Boolean formula that represents the conjunction of introduced arguments, together with the positions of agents toward this formula. The general layout of such a formula is given in (1). Debates are initialised without any arguments, and so the first debate stage is an empty formula.

$$\underbrace{((p_a \land ... \land p_b) \implies p_c)}_{\text{Argument 1}} \quad \land \quad \underbrace{((p_d \land ... \land p_e) \implies p_f)}_{\text{Argument 2}} \quad \land \quad ... \tag{1}$$

4.3 Agents, their belief systems and distances between them

Agents are fully aware of all available propositions, but they are not aware of propositions that are not yet introduced into the forum. Agents can either accept a proposition, reject it, or decide not to assign a truth value to that proposition.[5] Agents

[5] The state of refraining from passing judgement on a proposition is a first approximation to judgement suspension as a state in its own right (following Friedman, 2013). Research on the different kinds of suspension (Zinke, 2021) and how to implement them in agent-based models is still ongoing (Schuster, 2022).

assign one of these states to each proposition they are aware of, where acceptance is marked by "True", rejection by "False" and withholding by "None". An agent's belief system, or *position*, is represented by a mapping from the atomic propositions to these three states, e.g.:

$$\begin{aligned} p_1 &\rightarrow \text{True} \\ p_2 &\rightarrow \text{False} \\ p_3 &\rightarrow \text{None} \\ &\vdots \\ p_n &\rightarrow \dots \end{aligned} \tag{2}$$

This representation differs from that in many models in the philosophical literature, in which agents' beliefs are modelled with respect to a single proposition and then numerically in [0, 1] (O'Connor & Weatherall, 2018; Olsson, 2013; Pallavicini et al., 2021; Hegselmann & Krause, 2009; Zollman, 2007) or as a real number in $[-n, n]$, where n is determined by number and weight of reasons (Singer et al., 2019). While deliberation in these models is focussed on a single proposition, agents in the present model exchange arguments about all items in the sentence pool. But these propositions remain abstract representations of actual statements. The model can not account for the difference in, say, debates in politics compared to those in science, or between constructive deliberation and pointless exchange. It can be interpreted as a model for meetings of the Aristotelian Society – but also for Monty Python's Royal Society for Putting Things on Top of Other Things.

Agents are not allowed to adopt just any belief system, but are required to assign truth values in compliance with three rationality criteria: coherence, closedness and responsiveness. A belief system is coherent if it is free of contradictions. Agents are responsive in two ways, first by assigning a truth value to all propositions currently under discussion, and by accounting for all presented arguments in their choice of beliefs – irrespective of who introduced them. Agents account for an argument by selecting beliefs that allow the argument to be valid. This validity condition can be understood in logical terms as well: at each debate stage, a belief system must correspond to one of the interpretations of the Boolean formula that satisfy it. Closedness compels the agents to follow the arguments where they lead them: if an agent accepts all premises of an argument, it must accept the conclusion as well. There is a dynamic aspect to these rationality criteria. Coherence, closedness and responsiveness depend not only on agents' beliefs but also on the current debate stage: beliefs can be coherent and/or closed at stage τ_i, but become incoherent and/or not closed at τ_{i+1}, as the agents have to respond to newly presented arguments. Section 4.4 describes how agents respond to them and react in case of rationality violations.

These criteria describe agents' behaviour toward the forum – but there are also constraints on how the agents shape this forum in argument introductions, described in Sect. 4.4.1.

Differences between agents are measured as the distance between their belief systems, which is also the base value for measuring issue polarisation between them. The distance between two positions P_1, P_2 is measured by the edit distance $\text{ED}(P_1, P_2)$, which is defined as the minimal number of operations that have to be performed in order to change P_1 into P_2. ED allows three operations: switching a truth value assignment (from True to False or vice versa), adding a truth value to a position (i.e. changing from None to True or False) and removing a truth value from a position (i.e. changing from True or False to None).[6] As an example, consider how the edit distance between the two positions in (3) is 3:

$$\begin{array}{ll} \text{Position 1} & \text{Position 2} \\ p_1 \rightarrow \text{True} & p_1 \rightarrow \text{True} \\ p_2 \rightarrow \text{None} & p_2 \rightarrow \text{False} \\ p_3 \rightarrow \text{True} & p_3 \rightarrow \text{False} \\ p_4 \rightarrow \text{False} & p_4 \rightarrow \text{False} \\ p_5 \rightarrow \text{True} & p_5 \rightarrow \text{None} \end{array} \tag{3}$$

In (3), to obtain Position 2 from Position 1, p_2 has to be added, the truth value assignment of p_3 switched and the assignment of p_5 removed. The ED is symmetric as long as all operations are equally costly, or are weighted uniformly. As the necessary actions can differ for inverse operations, that is not true in the general case.[7]

The edit distance is not robustly meaningful in absolute terms. It makes a lot of difference that $\text{ED}(x, y) = 5$ whether the two positions x and y debate 10 or 100 propositions. *Normalising* the distance gives it a more universal meaning:

$$\frac{\text{ED}(x, y)}{|x \cup y|},$$

where $|x \cup y|$ is the size of the union of the positions' ranges. If (but only if) all operations are equally costly, the normalised edit distance is a variant of the widely used Jaccard distance, and reduces to the normalised Hamming distance.

4.4 Events

A computer simulation progresses by scheduled events. Simulations on the present model progress by either argument introduction or proposition pool expansion. Both are always followed by a position updating event, but they usually have a different chance of occurring (9:1 by default). Figure 1 gives an overview of a simulation run, the elements of which are explained in the following subsections.

4.4.1 Argument introduction

The argument introduction procedure selects two random agents from the population, of which the first is called *source* and the second *target*. All agents have the same probability of communicating with each other, irrespective of how much they agree. The whole population is immediately aware of the communication between

[6] These three operations are equivalent to Gärdenfors's "kinds of belief changes" (1992, 3).

[7] If and to what extend these operations should be weighted differently is a worthwhile question, but one that lies outside the scope of the present paper.

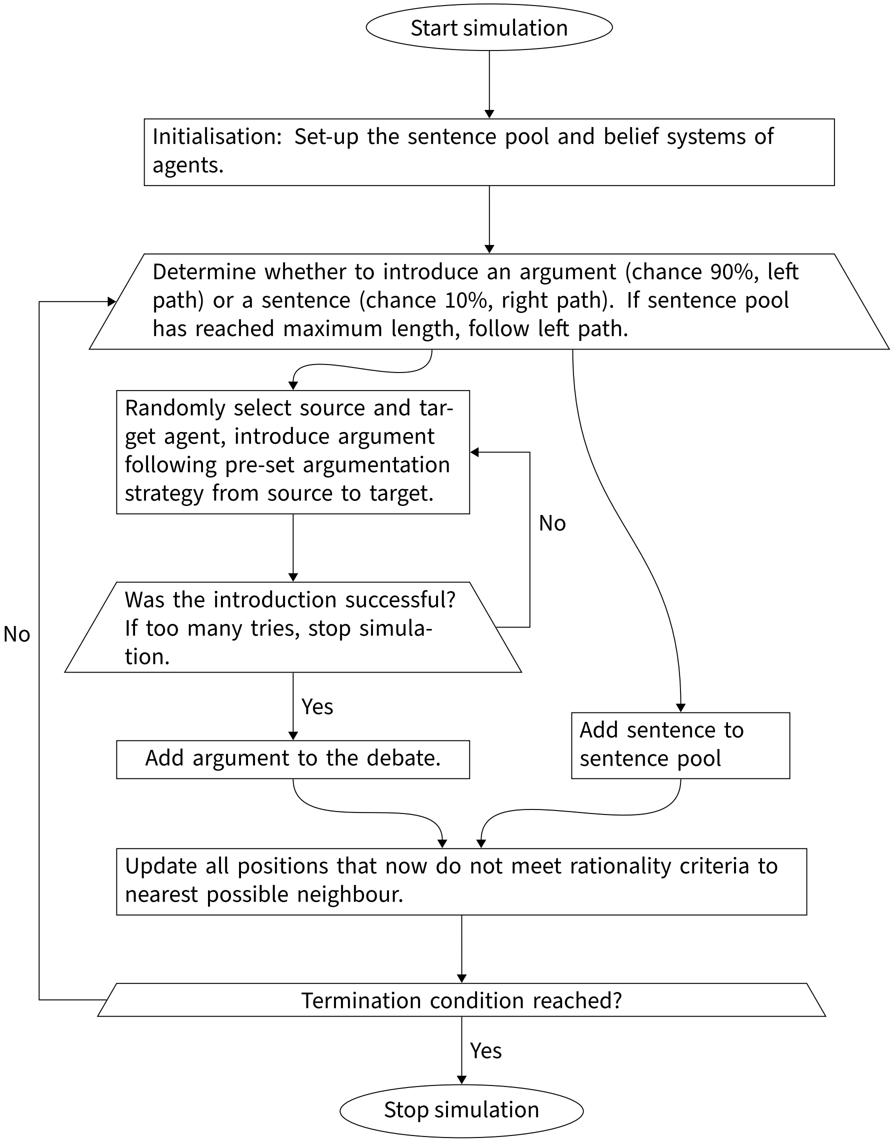
Start simulation
Initialisation: Set-up the sentence pool and belief systems of agents.
Determine whether to introduce an argument (chance 90%, left path) or a sentence (chance 10%, right path). If sentence pool has reached maximum length, follow left path.
Randomly select source and target agent, introduce argument following pre-set argumentation strategy from source to target.
No
Was the introduction successful? If too many tries, stop simulation.
No
Yes
Add argument to the debate.
Add sentence to sentence pool
Update all positions that now do not meet rationality criteria to nearest possible neighbour.
Termination condition reached?
Yes
Stop simulation

Fig. 1 Overview of a simulation run. Rectangles contain events and decisions are marked by trapeziums

source and target. The source first selects a conclusion and then premises from the pool of propositions, where the number of premises is a user-configurable random choice that defaults to 2 or 3 sentences. Following an idea by Betz, Chekan, and Mchedlidze (2021, §3), the conclusions are selected in such a way that the debate grows like a tree, with arguments that contain key statements as conclusions at the root of the resulting argument map (see Fig. 2). The premises of arguments that lead to key statements are then selected as conclusions on the second tier and the same

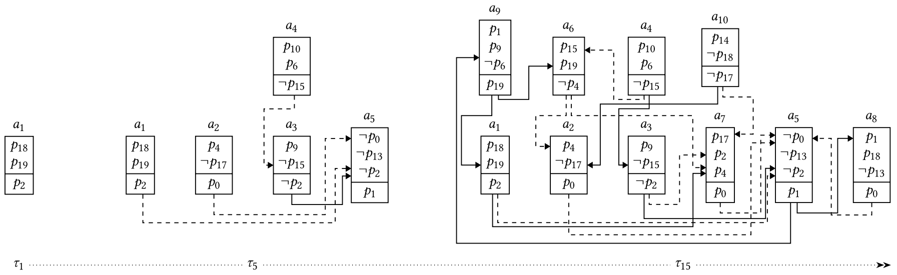

Fig. 2 Tree-like debate growth with key statements p_0, p_1 and p_2, showing three debate stages τ_1, τ_5 and τ_{15}

pattern repeats. Without such an hierarchical ordering, the argument map would take the shape of a random graph without discernible key issues.

There are two fundamental relations between arguments in the theory of dialectical structures, defeat and support. In Fig. 2, arguments are resembled by a two-part rectangle. The upper part collects the premises and the conclusion is in the lower part. If the conclusion of argument *a* is equivalent to one of the premises of argument *b*, *a supports b*. However, if the conclusion of *a* is equivalent to *the negation of b*, then *a defeats b*. Per convention, support is visualised with solid and defeat with dashed edges. The support and defeat relations are automatically determined from the introduced arguments and are not considered by the agents in advance.

To add an argument to the debate, the source selects premises and a conclusion that meet the criteria of the argumentation strategy the agent pursues. This can be one of five argumentation strategies (from Betz, 2013, pp. 93–94):

Attack: The source picks premises that it accepts and a conclusion that the target rejects or suspends on.

Fortify: The source selects both premises and a conclusion that it accepts. The position of the target is not considered.

Convert: The source selects premises that the target accepts and a conclusion that the source accepts.

Undercut: The source picks premises that the target accepts and a conclusion that the target does not accept (i.e. rejects or suspends on).

Any: One of the other strategies is followed randomly at each step.

Each of these abstract strategies represents a variety of actual argumentative behaviour. The fortify strategy, for example, captures how Cartwright (2013) thinks about finding evidential support for a hypothesis: the agent selects one or more premises about evidence, possibly together with auxiliary premises on general scientific procedures or principles, to support a conclusion it accepts. Change the conclusion to the negation of a hypothesis that the target accepts and an attack argument disconfirming the target's beliefs would emerge.

Beyond the strategy-dependent criteria, the source also ensures that the constructed argument meets two additional criteria to ensure that the other agents can respond rationally. The first constraint is purely internal to the argument. The premises in conjunction with the conclusion need to be free of contradictions and redundancies: the conclusion nor its negation are used as a premise. The second requirement is external to the new argument and concerns validity. After its introduction, at least one belief system respecting the validity of all arguments needs to remain for the agents to adopt – or, in logical terms, the debate's Boolean formula needs to remain satisfiable. Arguments often render previously held beliefs inadmissible – but in this model they must allow agents to revise their beliefs in accordance with the rationality criteria described in Sect. 4.3.

When the argument is introduced, the combination of premises is added to a list of used combinations and will not be introduced as part of another argument. In case the introduction fails to meet at least one of the conditions above, the process is repeated, with a fresh pair of agents, until a user-specified number of maximal tries is reached. In the exceedingly rare case that none of these tries yields an admissible argument, the simulation is terminated.

4.4.2 Proposition pool expansion

In this event, a proposition and its negation are entered into the debate forum – unless the maximum number of atomic propositions is reached, in which case this event has no effect. This event is reminiscent to (1) the case in *Twelve Angry Men* where juror 8 discovers that the murder weapon, a switch knife with an eye-catching design, is indeed readily available in shops near the scene, or (2) the case in which some agents remember that the two witnesses are not that reliable as the prosecution would have it. An introduced proposition is available for all subsequent argument introductions.

4.4.3 Position updating

Agents revise their beliefs after argument introduction and proposition pool expansion to uphold an epistemically rational outlook on the debate.[8]

The updating following the introduction of a new proposition is rather simple: every agent randomly assigns one of True, False, or None to the new proposition. Agents do not consider who introduced the sentence or for which conclusion the sentence might be used as a premise. Since newly introduced propositions are not yet used in any argument, agents do not risk violating rationality criteria through random assignment. Random assignment accounts for the fact that agents might have formed beliefs about new propositions in previous observations or deliberations before joining the current debate.

Position updating following argument introduction is a more complex process. Agents verify that their currently held beliefs are coherent, closed and allow all presented arguments to be valid. Any agent that does not hold such a position moves

[8] Belief revision is only triggered in these two ways to isolate the polarising effects of argumentation. The model in general would be adjustable to other forms of revision and to other forms of agent interaction.

to a new coherent, closed and responsive position with minimal edit distance to its current one, meaning a position that requires minimal belief revision. There is some motivation to think minimal adaptation is a rational move in the literature,[9] but it is also motivated in light of the present model. As agents respond to rationality violations induced by argument introduction, they have to consider that (1) there were arguments before the current one which motivated their current position and (2) there will be further constraints from arguments in the future. A move to *any* position compatible to rationality criteria could drastically change this agent's belief system. This would give the current argument immensely disproportionate influence, whereas responding through minimal adaptation does not give an argument too much preference over other arguments.

Often there are multiple ways to repair inadmissible beliefs that require the same number of belief revisions, implying identical edit distance. In this case, the agents have no preference what to do but decide randomly. Occasionally, agents can even be moved to suspension, rejection, or acceptance of all propositions under discussion.

5 Simulation procedure and results

5.1 Measurements

The model runs are interpreted with two measures of issue polarisation from Bramson et al. (2017, §2.7–2.8): group divergence and group consensus. Both measures assume that the population has been partitioned into clusters, or groups. Based on this partition, group divergence tracks how much more similar the belief systems among members of the same group are compared to agents in other groups. Group consensus measures how alike the groups are internally. Rising group divergence accompanied by rising consensus captures an intuitive understanding of polarisation very well: when this happens, groups become both more internally alike and externally alien.[10]

[9] For example, Singer et al. (2019) defend coherence-mindedness as rational concerning their agent-based model, but the rationality of minimal changes is also defended in texts about belief revision. Quine and Ullian (1978, 66–67) write (their emphasis): "Virtue I is *conservatism*. In order to explain the happenings that we are inventing it to explain, the hypothesis may have to conflict with some of our previous beliefs; but the fewer the better. Acceptance of a hypothesis is of course like acceptance of any belief in that it demands rejection of whatever conflicts with it. The less rejection of prior beliefs required, the more plausible the hypothesis – other things being equal."

The behaviour of the agents in this model is precisely that of Gärdenfors's *coherence theory* (Gärdenfors 1992, 8, his emphasis): "[A]ccording to the coherence theory, the objectives are, first, to maintain *consistency* in the revised epistemic state and, second, to make *minimal changes* of the old state that guarantee sufficient overall coherence."

[10] This is the conjunction of features 1 and 2 in Esteban and Ray (1994, 824). Their 3rd conceptual feature of polarisation, presence of a small number of significantly sized groups, is also realised by the clustering reported below. The clustering algorithms return between 2–4 clusters on the population of 12 agents.

Bramson et al.'s measures are defined on agents with single beliefs in the [0, 1] range. But there is no straightforward way to map the multi-dimensional belief systems in the theory of dialectical structures to the one-dimensional [0, 1] range. The measures can be adapted to the present model by operating on the differences between agents instead. These differences are given by the normalised edit distance and take values in the [0, 1] range. The obtained values for group divergence (Definition 1) and group consensus (Definition 2) lie in this interval as well.

Definition 1 Group divergence, based on Bramson et al. (2017, §2.7). Let A_τ be the population of agents at debate stage τ, represented by their positions. Let δ be the normalised edit distance. For a position x_i, $G(x_i)$ is the set of positions in the same group, while $G^*(x_i)$ are the out-group positions determined by a community structuring algorithm. Note that $|\cdot|$ denotes either the cardinality of a set or the absolute value of a distance.

$$\text{divergence}(\tau) := \frac{1}{|A_\tau|}\sum_{i}^{|A_\tau|}\left|\frac{\sum_{j\in G(x_i)}\delta(x_i,x_j)}{|G(x_i)|}-\frac{\sum_{k\in G^*(x_i)}\delta(x_i,x_k)}{|G^*(x_i)|}\right|$$

Note: The egocentric "me" in the measure runs on index i. Its neighbours run on index j and its strangers on k.

Definition 2 Group consensus, based on Bramson et al. (2017, §2.8). Let δ be the normalised edit distance and G the clustering of the population at a debate stage with individual clusters g. The expression $\binom{g}{2}$ is understood to denote the set of pairs in g. The debate's consensus is then given as:

$$\text{consensus}(\tau) := 1 - \frac{1}{|G|}\sum_{g=1}^{|G|}\frac{1}{\left|\binom{g}{2}\right|}\sum_{(x,y)\in\binom{g}{2}}\delta(x,y)$$

The partitioning into groups, or simply "clustering", required to calculate these values is obtained through two state-of-the-art community structuring algorithms for social networks, Leiden (Traag et al., 2019) and affinity propagation (Frey and Dueck, 2007).[11] Using multiple algorithms is one strategy to verify that the obtained clusterings are reliable.

[11] The clustering algorithms are run on distance matrices M with $m_{i,j} = \mathrm{ED}(i,j)/|i \cup j|$. This matrix is multiplied by an exponential scalar and filtered for values below 0.2 in order to generate sparsely populated matrices. This is necessary because community structuring algorithms are designed for social networks, in which most agents have relations to only a few other agents. Scaling and filtering are only applied to determine clusters. Divergence and consensus are then calculated based on the raw distances between agents' positions.

5.2 Simulation parameters

Simulations on the model are variable in initialisation and termination parameters, and they can be configured using the computational notebooks from the supplementary materials. For the initialisation, the number of agents and their initial belief systems, the initial sentence pool extension, the number of additional sentences for introduction, the number of premises per argument and the argumentation strategy shared by all agents can be set.

The results presented below were all obtained on populations of 12 agents pursuing the same argumentation strategy throughout a model run. The model runs have different initial sentence pools and belief distributions. Beliefs are assigned randomly in Sect. 5.3, antecedently polarised in Sect. 5.4 and initialised with 80% agreement in Sect. 5.5. Sections 5.4 and 5.5 present variations in which the entire sentence pool is known from the start and no further sentences are introduced in the course of a debate.

Although runs of the model could be terminated by specifying a maximum number of argument introductions, termination is here controlled by *inferential density*. This concept requires a little bit of a review. Agents are subjected to rationality constraints concerning their argumentative behaviour: they can only introduce arguments that cohere with the previously introduced arguments to the debate (in the sense that belief systems exist that accept the validity of all presented arguments), and they can only update their belief systems in a way that maintains validity of all presented arguments. Newly introduced arguments can render previously legitimate positions indefensible, but arguments differ in their impact regarding the number of positions they eliminate. Eventually, only one position is available for agents' belief systems, and no further arguments can be introduced. This ideal point is marked by an inferential density of $D = 1$. The initial stage at which all possible combinations of beliefs are admissible is marked by $D = 0$. The number of introduced arguments to reach $D = 1$ would not be a reliable indicator of simulation progress as it can differ significantly between model runs. Inferential density (defined in Betz, 2013, §2.5) is used as the normalised measure of simulation progress instead. It accounts for the number of positions rendered incoherent so far, or the freedom of movement that agents have in position updating.

Following Betz (2013, 95), debates are terminated at $D = 0.8$ by default. Depending on argumentation strategy, simulations take between 70 and 110 argument introductions on average to reach $D = 0.8$. Section 5.6 varies this termination condition by looking at the dynamics beyond $D = 0.8$.

5.3 Results from randomly allocated belief systems

Figures 3 and 4 show polarisation dynamics in 1,000 model runs per argumentation strategy with randomly initialised belief systems. Random initialisation means that each of the 12 agents assigns a random value of True, False or None to each proposition known to the initial forum. The entirely random beliefs account for the fact that agents meet after previously collecting evidence and engaging in other deliberations

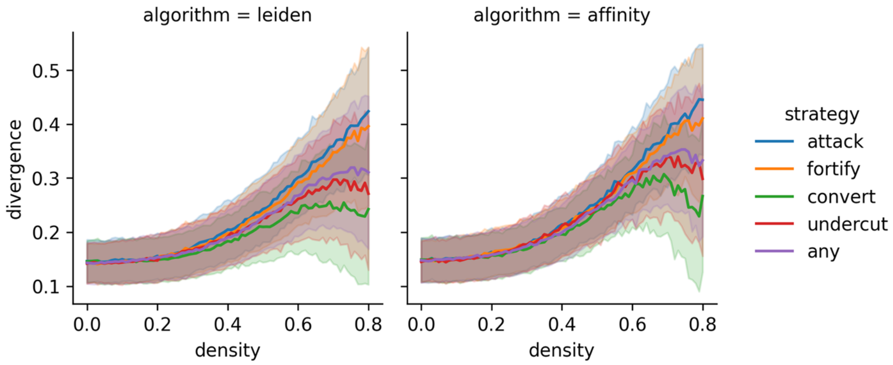

Fig. 3 Group divergence dynamics from two clustering algorithms under starting condition of low polarisation (completely random position assignment). The mean of 1000 runs per strategy is shown in the line plot, and the data's variation of ±1 standard deviation is plotted in the adjacent shaded area

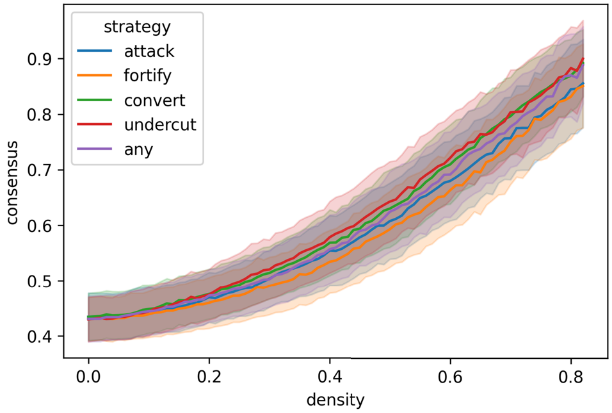

Fig. 4 Group consensus dynamics from Leiden clusterings under starting condition of low polarisation (completely random position assignment). The mean from 1000 model runs per strategy is plotted as a line, and variation of ±1 standard deviation is indicated by the adjacent area

before the modelled debate commences, but several robustness analyses in Sect. 5.4 and Sect. 5.5 verify the results for other belief initialisations. In this base experiment, the debate forum is initialised with a sentence pool of 15 propositions. On average, agents thus accept, reject and suspend on five propositions. Five more propositions were introduced in the course of the debate, resulting in a sentence pool size of 20. This limit is determined by the computational capabilities of the current software implementation and run-time on a state-of-the-art HPC, not by the model itself.

The data indicate that argumentation can be a driver of issue polarisation dynamics among rational agents. Polarisation here is understood as the increasing formation of internally coherent but externally divergent opinion clusters. As agents take on random positions initially, they have about the same distance to

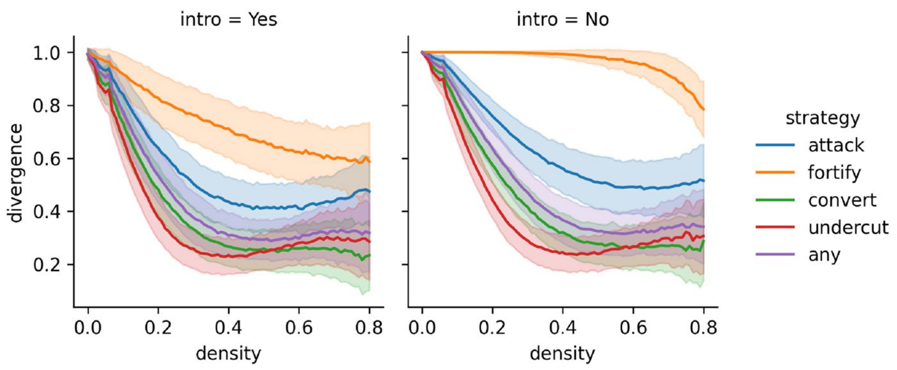

Fig. 5 Group divergence dynamics from 1000 runs per strategy following Leiden clusterings with groups antecedently configured to have maximum polarisation. The left, but not the right plot, shows experiments with proposition pool expansion. Means and standard deviation are indicated as before

most other agents. This implies low group divergence and medium consensus. The group-internal agreement and disagreement with out-group agents is low at this point, or, in other words, agents' beliefs are not well characterised by belonging to an opinion cluster. Beginning at these levels, the introduction of arguments is accompanied by a rise in divergence and consensus. This implies that agents form increasingly tight opinion-based groups and that these internally coherent groups grow farther away from other, likewise coherent groups.

But it also appears to matter *how* the agents argue with each other. The differing effects of argumentation strategies can be divided into two cases: simulations on the attack and fortify strategies exhibit a significantly higher group divergence compared to convert and undercut simulations, while the any strategy incorporates effects from all strategies. This result is interesting because it cuts along another division: in undercut and convert arguments, the source agent takes the target position into consideration for premise selection – "allocentrically", as it were. In arguments following the attack and fortify strategy, however, the introducing agent only considers its own position in premise selection, thus showing egocentric behaviour by the same standard. This observation is worth keeping in mind for the discussion (Sect. 6).

5.4 Results from antecedently opposed beliefs

Agents starting a debate with an entirely random initial belief distribution might be a rare encounter in the real world. A more common assumption is that agents enter debates belonging to different groups. Examples include those that maintain a defendant's guilt versus those that hold the defendant innocent, or proponents of different scientific theories.

The results in Fig. 5 provide robustness analyses for agents antecedently clustered into perfect tri-polarisation. In the first analysis (left), the debate started with a sentence pool of 15 (with positions as in (4)), which eventually expanded to 20. Four agents were assigned to each group.

$$\begin{array}{lll} \text{Group 1} & \text{Group 2} & \text{Group 3} \\ p_0, ..., p_4 \rightarrow \text{True} & p_0, ..., p_4 \rightarrow \text{False} & p_0, ..., p_4 \rightarrow \text{None} \\ p_5, ..., p_9 \rightarrow \text{False} & p_5, ..., p_9 \rightarrow \text{None} & p_5, ..., p_9 \rightarrow \text{True} \\ p_{10}, ..., p_{14} \rightarrow \text{None} & p_{10}, ..., p_{14} \rightarrow \text{True} & p_{10}, ..., p_{14} \rightarrow \text{False} \end{array} \tag{4}$$

In the previous experiment, the sentence pool expanded by a third of its original size. This feature is absent in the second scenario (Fig. 5, right), where 21 propositions in the sentence pool were known to the agents initially (5) and no new propositions were added in the course of the debate. This isolates the effect of sentence introduction and the unbiased evaluation of new sentences.

$$\begin{array}{lll} \text{Group 1} & \text{Group 2} & \text{Group 3} \\ p_0, ..., p_6 \rightarrow \text{True} & p_0, ..., p_6 \rightarrow \text{False} & p_0, ..., p_6 \rightarrow \text{None} \\ p_7, ..., p_{13} \rightarrow \text{False} & p_7, ..., p_{13} \rightarrow \text{None} & p_7, ..., p_{13} \rightarrow \text{True} \\ p_{14}, ..., p_{20} \rightarrow \text{None} & p_{14}, ..., p_{20} \rightarrow \text{True} & p_{14}, ..., p_{20} \rightarrow \text{False} \end{array} \tag{5}$$

The results indicate that argumentation can also drive depolarisation in debates. But this effect differs substantially between argumentation strategies as well. The allocentric strategies, convert and undercut, induced the lowest levels of polarisation in the previous experiment and now induce the strongest effect of depolarisation. In both cases, they terminate in similar polarisation values.

The egocentric strategies drive a much smaller effect of depolarisation and can terminate in higher values than in the previous experiment. These strategies seem also most affected by the unbiased evaluation of new propositions. Fortify debates in particular remain in near-perfect tri-polarisation for a long time when no sentences are introduced. It appears that giving agents the opportunity to evaluate newly introduced propositions without a bias enables them to find common ground with other agents in these newly acquired beliefs.

5.5 Initially agreeing agents and the effects of a fully known sentence pool

Another way to initialise a population of agents is to have them agree on most sentences under discussion. Just as random initialisation, this is a case of low initial polarisation. Figure 6 (right) shows results for a robustness analysis in which all agents share randomly allocated beliefs with 80% agreement and do not introduce further sentences.

Figure 6 (left) shows a robustness analysis with random initialisation, resulting in low initial agreement and polarisation, but where the complete sentence pool is known to the agents from the start. This further isolates the effect of sentence introduction.

In the initially polarised populations in Sect. 5.4, the introduction of additional sentences influenced the results and particularly affected the fortify strategy. In

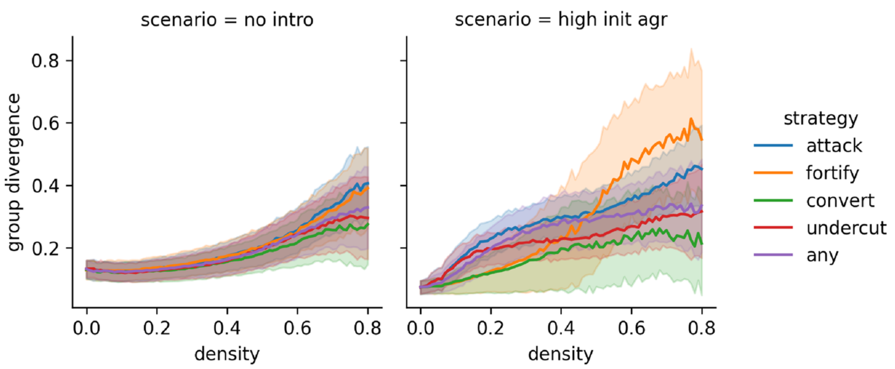

Fig. 6 Mean group divergence determined through Leiden clusterings for 1000 simulation runs per strategy. These are variations of the base experiment in which 20 sentences are known initially and no new sentences are introduced (left) and with initially highly agreeing (80%) populations and a fully known sentence pool (right). The shaded area displays values ±1 standard deviations away from the mean, which is indicated by lines

initially random beliefs, however, no noticeable deviance from the original results can be observed (Fig. 6, left, note the different *y* axis scale compared to Fig. 3).

Polarisation dynamics are observable even if agents initially agree on most sentences (Fig. 6, right) – in fact, the observed fortify values in this scenario are among the highest in this study. The fortify strategy is not only able to maintain high polarisation for a considerable time in polarised groups (Fig. 5), it also appears able to break up agreement among highly agreeing agents. Beyond the polarising fortify strategy, Fig. 6 also indicates a higher variation compared to scenarios with initially low agreement. Particularly noteworthy is the very low polarisation occasionally induced by the convert strategy.

5.6 Continuing debates to maximum inferential density

Debates can run for longer then until the termination density of $D = 0.8$, although it is questionable whether these debate stages correspond to any situation observable in the real world. In the ideal point of $D = 1$, inferential obligations would be so tight that rational agents have no choice but to settle on one remaining position. This leads to perfect agreement, which implies absence of issue polarisation.

Argumentation strategies differ in how agents approach this ideal point. Figure 7 illustrates this difference through randomly sampled model runs for convert and fortify. The *x* axis is not normalised by density in these plots, but by distance from the debate stage at which $D = 1$ is reached. In the fortify strategy, agents frequently uphold medium and high levels of group divergence until it becomes impossible for rational agents to disagree. Convert runs approach the ideal point much more gently. Shortly before reaching maximum density, the majority of runs have already reached very low divergence values.

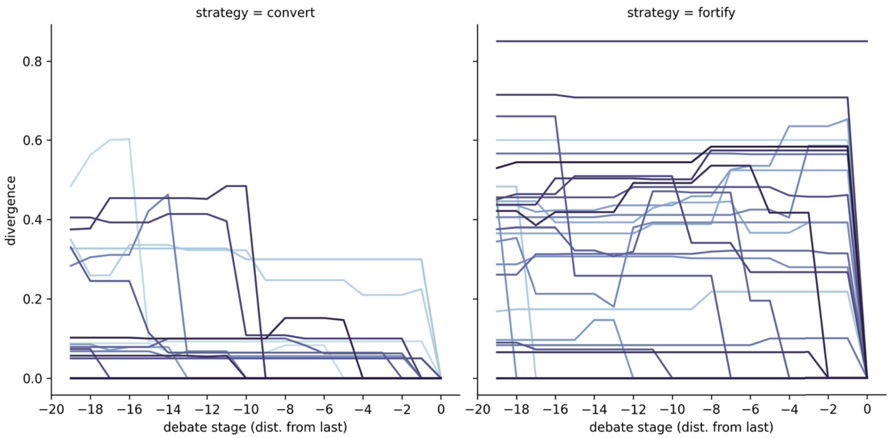

Fig. 7 Group divergence following Leiden clusterings in 40 randomly sampled convert and fortify runs that continue until maximum inferential density. The single fortify run that does not collapse is an example where termination occurs before the maximum density is reached, presumably due to unsuccessful argument introduction

6 Philosophical implications of the simulation results

Simulations on the present model reveal that argumentative behaviour can increase issue polarisation among artificial agents – even if their behaviour is constrained by epistemic rationality demands. This result gives further support to the possibility of rising polarisation under condition of epistemic rationality. In comparison to earlier approaches, it proves unnecessary to limit how agents can engage with their epistemic surroundings, such as through memory restrictions or by inhibiting communication through mistrust.

Which conclusions should we draw from the fact that computational models indicate plausible ways in which agents polarise even under condition of epistemic rationality? What consequences should we draw, in particular, regarding suggestions to intervene in polarisation dynamics? Are we justified in issuing negative evaluations or intervention recommendations for rationally induced polarisation dynamics?

It is important to remember that agents do not polarise affectively in the present model. While they polarised on their issue positions, they did not cease deliberative interaction even with the most remote of beliefs. Ceasing communication would be expectable if the agents were to polarise affectively. But when communication is upheld, epistemic communities can indeed operate under and recover from states of high issue polarisation. The geological community of the early 20th century moved from polarisation to agreement in light of more convincing data. And some of the group-inducing debates in philosophy have been going on for quite a while and it is not obvious that philosophy conferences have turned aggressive or epistemically less productive as a result (even if a discussion turns aggressive or unproductive, it is not immediately obvious that issue polarisation is the cause). Some even claim that issue polarisation can facilitate fruitful outcomes of discussions.[12] So if issue polarisation can rise among epistemically rational agents, is not an uncommon sight and can be mitigated by responsiveness to new and better data – then we should not necessarily consider it a *bad thing* that requires intervention.

The results go beyond the mere possibility of epistemically rational issue polarisation. They elucidate the influence of argumentation on this process and the differential effects that argumentation strategies have. These results underpin the relevance and impact of argumentation in social-epistemic processes. Our shared epistemic landscape is shaped through argumentation and particularly by *how* we argue with each other.

One might think that being critical toward others was particularly conducive to polarisation. Support for this view could be found in the results from the attack strategy, but the results also indicate that this is not always the case. The undercut strategy resembles a search for inconsistencies in the belief systems of others. This is a very critical approach, but the levels of polarisation induced by this strategy are low.

The results rather suggest that considering the beliefs of others at all is a more decisive factor compared to seeing these beliefs critically or favourably. When agents only work to fortify their own views and forgo engagement with others, polarisation rises substantially in case of initially low polarisation, and is particularly persistent in initially polarised groups. By comparison, only a minor polarisation effect in initially depolarised groups but a substantial depolarisation effect could be observed for initially polarised groups when agents select premises allocentrically, or in agreement with the beliefs of others. This underpins the productive effect that argumentation may have in conflict resolution – provided, in the present model, that agents remain in communication and engage with the views of others. When argumentation is interpreted as a general model for human reasoning (such as inspired by Mercier & Sperber, 2011), this indicates that reasoning allocentrically is conducive to preventing a rise in and reducing pre-existing polarisation.

In the fortify strategy, agents do not exhibit any behaviour toward others, critical or otherwise. What should we make of the fact that the strategy with the least social engagement leads to comparatively high polarisation values? Agents that pursue this strategy find more and more arguments supporting their currently held beliefs. Belief systems supported by many arguments, such as well-confirmed scientific theories, are the expectable outcome of this behaviour. This outcome is certainly desirable when applied to belief systems individually. And yet we must also recognise its polarising effect when applied by multiple agents with disagreeing beliefs. This raises a normative question: should we prioritise agreement and depolarisation and therefore compel agents with epistemic goals to engage in allocentric instead of egocentric reasoning? Or should we accept that high issue polarisation can be the consequence of rational and even virtuous individual behaviour?

[12] Popper (1976, 37) can be interpreted to agree with this claim when he writes that “fruitfulness in this sense will almost always depend on the original gap between the opinions of the participants in the discussion. The greater the gap, the more fruitful *can* the discussion be [...]” (his emphasis).

The insights into the epistemic impact of argumentation strategies also have methodological implications for the computational study of philosophical questions. They show that modelling epistemic behaviour with increased detail and realism can yield fruitful results – contrary to a sentiment previously expressed in the literature. Hegselmann and Krause (2009) defend a *low resolution approach* specifically with reference to more ambitious formal approaches that "so far did not deliver very much" (their fn. 2). A low resolution approach implies refraining from modelling "processes and actions of deliberative exchange" (2009, p. 131). The results obtained on the present model indicate that detailed models of deliberative exchange in general, and models built on the theory of dialectical structures in particular, can be philosophically productive – even though they do not adhere to the low resolution approach.

7 Conclusions

Simulations on agent-based models built on the theory of dialectical structures give further support to the thesis that debates among rational agents can polarise – in the specific sense of issue polarisation. Memory limits or trust networks need not be assumed to observe this phenomenon. Simulations run on the present model further show that there is a substantial difference in the impact of egocentric versus allocentric strategies in multi-agent reasoning. This extends earlier results on the influence of argumentation on consensus formation and the likelihood that groups of epistemic agents attain true beliefs through deliberation (Betz, 2013).

The influence of argumentation on polarisation dynamics underpins its role in understanding and evaluating epistemic processes, particularly in the social domain. The results also motivate reflection on how to judge occurrences of issue polarisation. Rather than seeking epistemic failure in a dynamic that is brought about rationally, we should underline the potential of eventual consensus if deliberative interaction is maintained – particularly when agents consider the views of others in their reasoning.

The results also have methodological appeal to computational philosophy projects. The perfectly valid and insightful results obtained from *low resolution approaches* (Hegselmann & Krause, 2009) do not imply that more ambitious models necessarily fail. Indeed, we should be looking forward to the new questions that computational philosophy will be able to tackle in more ambitious models.

Supplementary Information A repository at https://zenodo.org/record/6448599 contains the Python source code for an implementation of this model, together with Jupyter notebooks to run experiments and analyse the resulting data.

Acknowledgements For discussions of an earlier draft, I am thankful to Christian Seidel, Christian Straßer, Dunja Šešelja, Gregor Betz, Inga Bones, Julie Schweer, Michael Poznic, Sebastian Schmidt and Soong Yoo. Two anonymous reviewers pointed out significant potential for improvement in the original submission. Their time and work in conducting their constructive reviews is most appreciated. I also gratefully acknowledge support by the state of Baden-Württemberg through bwHPC, the state's high performance computing consortium.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Betz, G. (2009). Evaluating dialectical structures. *Journal of Philosophical Logic, 38*, 283–312.

Betz, G. (2013). Debate dynamics: How controversy improves our beliefs. Berlin: Springer. Retrieved from https://doi.org/10/d3cx

Betz, G., Chekan, V., & Mchedlidze, T. (2021). Heuristic algorithms for the approximation of Mutual Coherence. Retrieved from https://doi.org/10.48550/arXiv.2307.01639

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2022). Cross-country trends in affective polarization. *The Review of Economics and Statistics*. https://doi.org/10.1162/resta01160

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science, 84*(1), 115–159. https://doi.org/10.1086/688938

Burnstein, E., & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology, 13*(4), 315–332.

Cartwright, N. (2013). Evidence, argument and prediction. V. Karakostas & D. Dieks (Eds.), EPSA11: Perspectives and foundational problems in philosophy of science (pp. 3–17). Cham: Springer. Retrieved from https://doi.org/kmkh

Dutilh Novaes, C. (2021). Argument and argumentation. E.N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Fall 2021 ed.). Retrieved from https://plato.stanford.edu/archives/fall2021/entries/argument/

Esteban, J.-M., & Ray, D. (1994). On the measurement of polarization. *Econometria, 62*(4), 819–851.

Fogal, D., & Worsnip, A. (2021). Which reason? Which rationality? *Ergo*. https://doi.org/10.3998/ergo.1148

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972–976. https://doi.org/10.1126/science.1136800

Friedman, J. (2013). Suspended judgment. *Philosophical Studies, 162*, 162–181. https://doi.org/10.1007/s11098-011-9753-y

Gärdenfors, P. (1992). Belief revision: An introduction. P. Gärdenfors (Ed.), Belief revision (pp. 1–28). Cambridge, UK: Cambridge University Press.

Hallam, A. (1989). *Great geological controversies* (2nd ed.). Oxford University Press.

Hegselmann, R., & Krause, U. (2009). Deliberative exchange, truth, and cognitive division of labour: A low-resolution modeling approach. *Episteme, 6*(2), 130–144. https://doi.org/10.3366/E1742360009000604

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22*, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly, 76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology, 59*(1), 193–224. https://doi.org/10.1146/annurev.psych.59.103006.093615

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision making, 8*(4), 407–424. https://doi.org/10.2139/ssrn.2182588

Mäs, M., & Flache, A. (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

Mason, L. (2013). The rise of uncivil agreement: Issue versus behavioral polarization in the American electorate. *American Behavioral Scientist, 57*(1), 140–159. https://doi.org/10.1177/0002764212463363

Mason, L. (2015). "I disrespectfully agree'': The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science, 59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*, 57–111. https://doi.org/10.1017/S0140525X10000968

Mutz, D. C. (2002). Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review, 96*(1), 111–126. https://doi.org/10.1017/S0003055402004264

Myers, D. G. (1975). Discussion-induced attitude polarization. *Human Relations, 28*(8), 699–714. https://doi.org/10.1177/001872677502800802

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science, 8*, 855–875. https://doi.org/10.1007/s13194-018-0213-9

Olsson, E.J. (2013). A Bayesian simulation model of group deliberation and polarization. F. Zenker (Ed.), Bayesian argumentation: The practical side of probability (pp. 113–133). Dordrecht: Springer. Retrieved from https://doi.org/10/ggz2

Pallavicini, J., Hallsson, B., & Kappel, K. (2021). Polarization in groups of Bayesian agents. *Synthese, 198*, 1–55. https://doi.org/10.1007/s11229-018-01978-w

Pew Research Center (2014). Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise and everyday life. Retrieved from https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Pew Research Center (2017). The partisan divide on political values grows even wider. Retrieved from https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/

Popper, K. (1976). The myth of the framework. J.C. Pitt & M. Pera (Eds.), Rational changes in science: Essays on scientific reasoning (pp. 35–62). Dordrecht: D. Reidel.

Proietti, C., & Chiarella, D. (2021). Measuring bi-polarization with argument graphs. M. D'Agostino, F.A. D'Asaro, & C. Larese (Eds.), In Proceedings of the 5th workshop on advances in argumentation in Artificial Intelligence. Retrieved from http://ceur-ws.org/Vol-3086/paper6.pdf

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (2nd ed.). McGraw-Hill.

Schuster, D. (2022). Forms and norms of indecision in argumentation theory. Retrieved from https://doi.org/10.48550/arXiv.2203.02207 (Presented at the 15th international conference on deontic logic and normative systems, DEON 2020/2021)

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., & Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies, 176*(9), 2243–2267. https://doi.org/10.1007/s11098-018-1124-5

Sunstein, C. R. (2002). The law of group polarization. *The Journal of Political Philosophy, 10*(2), 175–195.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*, 5233.

Zinke, A. (2021). Rational suspension. *Theoria, 87*(5), 1050–1066. https://doi.org/10.1111/theo.12320

Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science, 74*(5), 574–587. https://doi.org/10.1086/525605