Polarisation, diversity, and dialectical structures

An argumentation-based approach to computational social epistemology

Zur Erlangung des akademischen Grades eines
DOKTORS DER PHILOSOPHIE (DR. PHIL.)

von der KIT-Fakultät für Geistes- und Sozialwissenschaften des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von
Felix Emanuel Kopecky

KIT-Dekan: Prof. Dr. Michael Mäs

1. Gutachter: Prof. Dr. Gregor Betz
2. Gutachterin: Prof. Dr. Dunja Šešelja

Tag der mündlichen Prüfung: 9. Juli 2025

Contents

Acknowledgements

The submission of this dissertation concludes a most intense period of learning and development, both academic and personal. I would like to express my deepest gratitude to Gregor Betz for having enabled this process through his supervision, guidance, advice, and co-operation.

This dissertation benefited considerably from a research stay at Ruhr-Universität Bochum in December 2022. I wish to thank Dunja Šešelja and Christian Straßer for their warm and kind hospitality and for the engaging discussions we shared during this stay.

I'd also like to express my gratitude to everyone who commented on earlier versions of the work presented in this dissertation, namely Anna Klassen, Christian Seidel, Christian Straßer, Christoph Merdes, David Lanius, Dunja Šešelja, Gregor Betz, Inga Bones, Julie Schweer, Klee Schöppl, Leon Assaad, Matteo Michelini, Maximilian Vossel, Michael Poznic, Rafael Fuchs, Sebastian Schmidt, Simon Derpmann, and Soong Yoo, as well as anonymous referees for the published papers. I'd like to thank the organisers of conferences and workshops in Berlin, Bochum, Helsinki, Münster, and Pisa for their welcoming hospitality and the opportunity to present the findings from this thesis, as well as the audiences for critical and productive discussions.

The Karlsruhe Institute of Technology and the state of Baden-Württemberg have supported this dissertation in several ways. I would like to explicitly thank the library for covering publication costs through the KIT Publication Fund as well as the staff at bwHPC, the high-performance computing consortium funded by the state of Baden-Württemberg. The simulation experiments that yielded the results of this dissertation were performed on bwUniCluster 2.0.

A travel stipend from the German Academic Exchange Service (DAAD) is gratefully acknowledged. This travel stipend received in 2023 covered the travel costs associated with a presentation in Pisa.

Publications

Peer-reviewed articles

The work presented in this dissertation is based on the following publications.

Kopecky, Felix (2022). Arguments as drivers of issue polarisation in debates among artificial agents. In: *Journal of Artificial Societies and Social Simulation* 25.1. DOI: 10.18564/jasss.4767.

Kopecky, Felix (2024). Argumentation-induced rational issue polarisation. In: *Philosophical Studies* 181.1, pp. 83–107. DOI: 10.1007/s11098-023-02059-6.

Kopecky, Felix & Gregor Betz (2025). Inconsistent belief aggregation in diverse and polarised groups. In: *Philosophy of Science* 92.1, pp. 40–58. DOI: 10.1017/psa.2024.29.

Chapters 3, 4, and 5 in this dissertation are slightly modified reproductions of the main parts of these publications. Material from these publications has also been reproduced verbatim in Chapters 1, 2, and 6 as well as in Appendix A.

Software

A Python package to study the theory of dialectical structures was implemented as part of this dissertation. Its source code is published as:

Kopecky, Felix (2021–2022). *taupy: A Python package to study the theory of dialectical structures*. DOI: 10.5281/zenodo.5067834.

The documentation for taupy, found in Appendix B, is reproduced with slight adjustments from the original web version at https://kopeckyf.github.io/taupy/.

Statement of authorship

The first two research articles and the software implementation listed above are the sole work of Felix Kopecky. The third paper listed above is collaborative work of Gregor Betz and Felix Kopecky. In terms of the CRediT contribution taxonomy, the work for Kopecky & Betz (2025) was shared as follows:

Conceptualisation: GB & FK – *Methodology*: GB & FK – *Software, validation, and formal analysis*: FK – *Investigation*: FK – *Data curation*: FK – *Writing (first draft)*: FK – *Writing (review & editing)*: GB & FK – *Visualisation*: FK.

Summary

This dissertation is about the study of belief polarisation and opinion diversity in agent-based, computational models. Simulations on these models reveal that beliefs of artificial agents can polarise through deliberation – even if all agents hold a productive mindset and adhere to standards of epistemic rationality. Different deliberative practices affect polarisation dynamics differently in these models. Polarisation dynamics are soothed as agents engage with the beliefs of others in their reasoning. The largest polarisation effect is observed in agents who continuously fortify their own beliefs.

A second result is that, when artificial agents vote on a collective response towards a body of arguments, the probability that their vote yields an inconsistent majority opinion is strongly associated with the group's opinion diversity. This improves our understanding of the difficulties that diverse groups can face in decision-making. When they can not use majority voting as part of their decision-making, they need to find alternative, likely more demanding, aggregation procedures to settle their difference of opinion.

These results are gathered from computational models based on the theory of dialectical structures. This theory describes deliberation through argument maps. Individual arguments in these maps are composed of premises and conclusions. The beliefs of agents in dialectical structure models are implemented as mappings from the discussed premises and conclusions to discrete truth values.

There are two kinds of dialectical structure models. In the first, agents iteratively add arguments to an evolving argument map, and they update their beliefs following the introductions by others. Both introduction of and response to arguments are constrained by factors that ensure epistemically rational behaviour. Beyond these constraints, agents compose new arguments according to pre-defined argumentation strategies. These strategies can be divided into allocentric and egocentric strategies, the two of which have fundamentally different effects on belief polarisation. The population polarises if agents only follow egocentric strategies, and the population depolarises if they argue allocentrically.

The second kind of dialectical structure model does not evolve through iterative additions of arguments. It instead synthesises an argument map according to an algorithm found in the literature on dialectical structures. An argument map generated in this way poses an epistemic decision problem. Agents are tasked with finding a belief system that jointly accepts the validity of all presented arguments. There are usually many beliefs that meet this criterion, but even more that do not. This decision problem becomes a problem in collective decision-making if agents with different, individually validity-respecting

beliefs need to settle on a single belief system to endorse collectively. In the model presented here, agents perform a sentence-wise majority vote to do so. These votes do not necessarily yield a consistent opinion. Analysis of computationally gathered data reveals that inconsistent aggregation is strongly associated with rising opinion diversity, but not with belief polarisation.

taupy, a new Python implementation for the study of dialectical structures, is implemented and released to the public as part of this dissertation. This implementation provides features that were not previously available. It implements the clustering of agents' beliefs with two state-of-the-art clustering algorithms, as well as measures of belief polarisation and opinion diversity, a majority voting mechanism, and other features necessary to observe the results presented here. A user guide is supplied in this thesis.

Dialectical structure models improve our understanding of dynamic aspects of concepts, constraints, and norms that we study in philosophy. This makes computational modelling such a fruitful approach to philosophy, particularly in social epistemology. In this thesis, the discovery of belief polarisation under condition of epistemic rationality reveals a dynamic property of our concept of rationality. Computational approaches are a useful addition to philosophical methods because such dynamic aspects would not be accessible from established methods of analysis.

This thesis pursues computational methods out of purely philosophical interest, but it relies on and contributes to inter-disciplinary efforts in several ways. Polarisation and diversity are understood as they are in other fields, and the agent-based modelling presented here is related to sociological models that have similarly suggested arguments as drivers of polarisation.

While this dissertation studies beliefs and their dynamics in artificial, epistemically rational agents, the results are relevant for our understanding of human deliberation and reasoning, too. As a matter of principle, the adherence to norms of epistemic rationality does not prevent the rise of belief polarisation. This means that occurrences of belief polarisation in humans can not automatically be taken as evidence of epistemic shortcomings in any individual. And the issue of inconsistent belief aggregation in groups with diverse opinions illustrates the difficulties groups with epistemic goals can face. This helps us understand the conditions and limitations of experts advising the public, particularly in exceptional situations involving high uncertainty and limited time. Results from this thesis suggest that we should indeed moderate our expectations towards these groups when exceptional circumstances hold.

Epistemically rational belief polarisation and inconsistent belief aggregation are instances of a broader phenomenon. These are cases in which the accumulation of individually rational capabilities does not yield an optimal outcome for the collective. By continuing the study of such intriguing phenomena, we can hope to gain insights into noteworthy conditions of human reasoning, rationality, and decision-making.

Zusammenfassung auf Deutsch

Diese Dissertation beschäftigt sich mit dem Studium von Meinungspolarisierung und Meinungsvielfalt in agentenbasierten Computermodellen. Simulationen dieser Modelle legen nahe, dass die Meinungen künstlicher Agenten allein durch den Austausch von Argumenten polarisieren können – und zwar auch dann, wenn alle Agenten ernsthaft zur Diskussion beitragen und den Normen der epistemischen Rationalität folgen. Unterschiedliche argumentative Praktiken haben aber einen unterschiedlichen Einfluss auf diesen Polarisierungseffekt. Dieser kann abgeschwächt werden, wenn Agenten die Meinungen anderer in ihre Überlegungen einbeziehen. Der stärkste Polarisierungseffekt zeigt sich in Agenten, die ihre eigene Meinung kontinuierlich festigen.

Ein zweites Resultat betrifft die Fähigkeit von künstlichen Agenten, durch Abstimmungen eine konsistente Mehrheitsmeinung zu bilden. Hier zeigt sich ein starker Zusammenhang zwischen Meinungsvielfalt und Inkonsistenzen in der Mehrheitsmeinung. Dieses Ergebnis deutet auf Schwierigkeiten hin, denen Gruppen mit vielfältigen Meinungen in Entscheidungsprozessen ausgesetzt sein können. Wenn eine Abstimmung keine konsistente Mehrheitsmeinung ergibt, muss die Entscheidung durch andere, wahrscheinlich anspruchsvollere, Wege herbeigeführt werden.

Die Theorie dialektischer Strukturen bildet den Rahmen für die Modellierungen dieser Dissertation. Diese Theorie beschreibt deliberative Prozesse durch Argumentkarten. In diesen Karten sind einzelne Argumente durch ihre Prämissen und Konklusionen dargestellt. Die Meinungen der Agenten sind Zuordnungen von in der Karte vorkommenden Sätzen zu diskreten Wahrheitswerten.

Es gibt zwei Arten von Computermodellen dialektischer Strukturen. In Modellen der ersten Art fügen Agenten kontinuierlich Argumente zur Argumentkarte hinzu und verändern ihre Meinung, wenn die Argumente anderer es gebieten. Sowohl diese Einführungen wie auch die Reaktionen auf Argumente sind dabei durch Standards epistemischer Rationalität reglementiert. Über diese Standards hinaus können Agenten unterschiedlichen Argumentationsstrategien folgen, welche auch substantiell verschiedene Polarisierungseffekte herbeiführen. Durch die Einführung von Argumenten aus egozentrischer Perspektive wird die Polarisierung verstärkt. Beziehen Agenten aber auch die Meinungen anderer ein, können die Meinungen sogar depolarisieren.

Modelle der zweiten Art kennen keine kontinuierliche Erweiterung ihrer Argumentkarte. Stattdessen wird durch einen aus der Literatur entnommenen Algorithmus eine Argumentkarte synthetisiert. Eine solche Argumentkarte kann als epistemisches Entscheidungsproblem verstanden werden, insofern sich Agenten eine Meinung bilden müssen, welche mit der Gültigkeit aller in

der Karte vorkommenden Argumente verträglich ist. Dieses Kriterium wird normalerweise von vielen Meinungen erfüllt, von einer noch größeren Anzahl aber verletzt. Dieses Entscheidungsproblem kann auch als kollektive Aufgabe verstanden werden, nämlich dann, wenn Agenten mit unterschiedlichen Meinungen eine für sie repräsentative Gruppenmeinung finden müssen. In dem hier verwendeten Modell erzielen die Agenten eine solche Gruppenmeinung per Abstimmung über jeden Satz. Diese Abstimmungen führen jedoch nicht immer zu Meinungen, welche mit der Gültigkeit aller Argumente verträglich sind. Es zeigt sich, dass das Risiko einer inkonsistenten Mehrheitsmeinung stark mit der Meinungsvielfalt in der Gruppe zusammenhängt – nicht aber mit deren Meinungspolarisierung.

Für diese Dissertation wurde eine neue Python-Implementierung für das Studium dialektischer Strukturen implementiert und veröffentlicht: *taupy*. Diese Implementierung bringt eine Reihe von Neuerung mit sich. Es ist nun möglich, Clusteringalgorithmen, Polarisierungsmaße und Diversitätsindizes in Modellen dialektischer Strukturen anzuwenden sowie Abstimmungen unter Agenten durchzuführen. Ein Handbuch ist dieser Dissertation beigefügt.

Computermodelle dialektischer Strukturen bereichern unser Wissen über dynamische Aspekte mancher philosophischer Begriffe und Normen. Dadurch erweist sich die Computermodellierung insgesamt als fruchtbarer Forschungsansatz in der Philosophie und insbesondere in der Sozialen Erkenntnistheorie. Die Entdeckung der Meinungspolarisierung unter Einhaltung von Rationalitätsnormen ist zugleich die Entdeckung einer dynamischen Eigenschaft unseres Rationalitätsbegriffs. Derartige dynamische Eigenschaften sind der bloßen Analyse von Begriffen verschlossen.

Diese Dissertation verfolgt die Computermodellierung ausschließlich aus philosophischem Interesse, doch bezieht sie sich an vielen Stellen auf die Erforschung von Polarisierung und Vielfalt in anderen Disziplinen. Eines der hier präsentierten Modelle könnte auch im Rahmen der computergestützten Soziologie interpretiert werden, wo der Austausch von Argumenten ebenfalls als möglicher Faktor von Polarisierung untersucht wird.

Obwohl diese Dissertation den Austausch von Argumenten nur in künstlichen, epistemisch stets rationalen Agenten untersucht, sind die Ergebnisse dennoch für das Verständnis von Meinungsdynamiken am Menschen aufschlussreich. So kann Meinungspolarisierung nicht immer allein durch die Einhaltung von Normen epistemischer Rationalität verhindert werden. Die Polarisierung unter Menschen kann also nicht automatisch als Beleg für irrationales Verhalten Einzelner herangezogen werden. Und das Phänomen der inkonsistenten Mehrheitsentscheidungen in Gruppen mit hoher Meinungsvielfalt verdeutlicht mögliche Problematiken der kollektiven Entscheidungsfindung. So können die Bedingungen und Beschränkungen besser verstanden werden, denen etwa die Beratung der Öffentlichkeit durch Expertengruppen unterliegen kann. Besonders unter außergewöhnlichen Umständen, wie besonders großen sachlichen Unwägbarkeiten und hohem Zeitdruck, erscheint eine zurückhaltende Erwartungshaltung gegenüber diesen Gruppen geboten.

Epistemisch rationale Meinungspolarisierung und inkonsistente Mehrheitsentscheidungen veranschaulichen auch ein umfassenderes Phänomen der menschlichen Rationalität. Es handelt sich jeweils um Fälle, in denen die Akkumulation von individuellen, rationalen Fähigkeiten nicht in einem optimalen Ergebnis für die Gemeinschaft kumuliert. Von einem weiteren Studium dieser erstaunlichen Phänomene kann man sich durchaus einen Erkenntnisfortschritt über relevante Aspekte der menschlichen Rationalität und Entscheidungsfindung versprechen.

Im Einklang mit der bestehenden Fachliteratur ist diese Dissertation auf Englisch verfasst.

Chapter 1

Introduction

1.1 Overview

Rationality is a blessing. In humans, it enables individuals to weigh reasons, form and revise beliefs, make decisions, and plan ahead. These rational capacities allow humans to develop powerful technologies and machines, distinguish lawful from wrong behaviour in courts, and uncover the mysteries of the skies. Rationality is self-correcting, too. It allows us to reflect on our biases and puts us in a position to correct them.

Humans do not only apply their rational abilities in isolation, but they convene in groups and powerful institutions to solve ever more complex problems with combined forces, such as in politics, engineering, or science. The success often achieved by such collaborations gives rise to an optimistic but plausible hypothesis: if agents collectively apply their faculties, their rational capabilities will cumulate and the outcome will greatly improve from the individual application, as the agents converge to a single state of mind that is uniquely suitable to solve the problem at hand.

Surprisingly, and somewhat worryingly, this convergence does not always materialise. The collective application of individually rational procedures can instead lead agents to diverge, even if all agents hold a productive mindset. These phenomena can be described as *independence phenomena* because they illustrate how the outcome of a collective application of rational capabilities can have qualities independent from their individual application.

This dissertation is about the study of independence phenomena in computational models. In the course of this dissertation, I will present two computational models in which artificial agents are rational individually but diverge from an output that can be considered optimal for the group. In the first model, the beliefs of agents can polarise as they exchange valid arguments and respond rationally to the arguments of others. In the second model, agents posed to collectively settle on a solution to a decision problem sometimes aggregate a group opinion that is irrational, even though they all hold individually rational beliefs. In the first model, the degree of polarisation is associated with distinct kinds of argumentative behaviour. Continuous unilateral belief fortification tends to polarise groups while considering the beliefs of others contributes to depolarisation. In the second model, the frequency of collectively irrational opinions is strongly associated with the group's opinion diversity, but not with its belief polarisation.

These results are gathered from computational models with idealising assumptions, but they can inform our understanding of very real phenomena. The possibility of belief polarisation under condition of epistemic rationality suggests that rising polarisation is not a reliable indicator of epistemic shortcomings. Polarisation can occur even if all participants engage in productive dialogue and adhere to norms of rationality. This might seem like a concerning result, but it does not necessarily suggest declining affections in agents. It is a worthwhile question for further research under which conditions humans maintain communication and mutual respect despite vastly different opinions. The results further suggest that different kinds of deliberative behaviour have different effects on polarisation. The way in which we consider the beliefs of others fundamentally shapes the opinion landscape we share with them.

Results from the second model indicate a strong relation between opinion diversity and the likelihood of inconsistent results in majority voting. Diverse groups that rely on voting as part of their decision-making are at a particular risk to aggregate inconsistent group opinions. I will argue that this improves our understanding of epistemic group problem solving under exceptional circumstances, such as a high degree of uncertainty and a limited amount of time. These findings help us understand the difficulties faced by expert groups and what we can reasonably expect from expert advice. Under some exceptional circumstances we should indeed moderate our expectations towards them.

This introduction covers the fundamental concepts underlying these results. First, a more detailed characterisation of independence phenomena follows in Section 1.2.

The concept of rationality is central to many parts of this dissertation. But what is rationality, anyway? In colloquial terms we sometimes say that a decision was "rational" and we mean not much more than that it was a good decision. And yet philosophers and psychologists have uncovered many more aspects about this complex concept. Section 1.3 contains a brief review of these many faces of rationality.

This dissertation is not only about modelling rational behaviour in artificial agents, but also about two dynamic properties of individuals and collectives, *diversity* and *polarisation*. Like the concept of rationality, these two concepts can be understood in a number of ways. The precise interpretation of these concepts is relevant to understand the conclusions that can be legitimately drawn from the computational models. I review characterisations of polarisation and diversity concepts in Sections 1.4 and 1.5, respectively. This informal discussion is supplemented by a more formal treatment in Chapter 2.

Social epistemology is a philosophical discipline that is occupied with collective rationality and independence phenomena. In the 21st century, formal and computational methods are emerging and productive approaches in social epistemology and neighbouring disciplines. This dissertation approaches issues in social epistemology from a somewhat lesser-known approach based on arguments and argumentation. In Section 1.6, I give reasons to consider this approach as particularly promising, and I close this introduction with a

brief reflection on the challenges, but also opportunities, that computational approaches to social epistemology provide (Section 1.7).

Beyond the introduction, Chapter 2 is an introduction to how computational models can be built with the framework underlying this work, the theory of dialectical structures. There I review its existing concepts, measures and methods, but I also introduce new ones to measure belief polarisation, opinion diversity, and information entropy. This chapter also introduces methods to cluster belief systems, a necessary ingredient for many polarisation and diversity measures. This chapter finally differentiates two different types of dialectical structure models. In the first, argument maps grow through the interaction of agents. In the second, argument maps are antecedently synthesised.

Chapters 3 and 4 are about belief polarisation in dialectical structure models and they establish the possibility of polarisation under condition of epistemic rationality. In these chapters I present the experimental simulation procedure and analyse the results. I also reflect on the now discovered possibility of rational polarisation and offer an explanation why we should not despair over this course of things.

Chapter 5 is about inconsistent belief aggregation in groups with varying degrees of opinion diversity and belief polarisation. In particular, the chapter explores the benefits and risks of diversity and polarisation in expert groups as they issue advice under permissive evidence, uncertainty, and a very limited amount of time.

The conclusion at the end (Chapter 6) summarises what can be learned from simulations on dialectical structure models.

1.2 Independence phenomena

Modern humans face many tasks that are so difficult and complex that they surpass the capabilities of any single individual. Although individuals can make most praiseworthy contributions through talent, learning and dedication, with rising task complexity their work is necessarily intertwined with the work of others. The success humans can hope to achieve in approaching cutting edge issues in science, modern engineering projects, or pressing political issues that involve diverse standpoints, depends in part on their ability to coordinate their commitment in groups.

This requirement to tackle issues in groups and to aggregate individual views to collective group opinions should not raise immediate concern. After all, there are now more helping hands, there is more talent in a group than in one person, and there is more learned experience the more people meet in a group. This optimism should be particularly warranted if all individuals in a group are rational and interested in solving the issues faced by the group.

But the optimal group outcome is not always achieved through addition of individually rational behaviour. Sometimes groups produce opinions that are sub-optimal relative to an optimal state regarding knowledge or belief. Things can even get worse: aggregating individually rational behaviour can lead to

outcomes that are outright irrational. These discrepancies between individual rationality and collective optima have previously been studied in the philosophical literature under different headings. Kitcher (1990, pp. 6–10) introduces their accurate description as "collective optimum–individual rationality discrepancies", and Mayo-Wilson, Zollman & Danks (2011) discuss them under the shorter heading of their "independence thesis". In the context of this dissertation, the name I wish to give for occurrences of such discrepancies is *independence phenomena*.

Independence phenomena have two ingredients: individually rational behaviour or beliefs on the one hand and a collective state that is considered optimal. Given the conditions of individual rationality that I review in Section 1.3, the first ingredient is not too difficult to characterise. There is, however, quite a bit of flexibility when it comes to determining what constitutes a collective optimum in an independence phenomenon. The state achieved by a group can be "optimal" in more than one way, and a group may achieve its optimum in one respect but not another. When describing an independence phenomenon, it is thus important to state exactly in which respect optimality should be achieved by the group.

For processes that involve the aggregation of opinions, the collective optimum could consist in the rationality of the opinion that the group collectively maintains. Independence phenomena of this sort are cases in which the rationality of the individuals does not guarantee that the group will hold a rational opinion as well. We can also say that rationality "is not conserved" through belief aggregation. Phenomena like this have long been known – they are already described by Nicolas de Condorcet in the late 18th century (de Condorcet 1785). But in the following I rely on the contemporary characterisation from List & Pettit (2002).

Let us consider this first example for an independence phenomenon in more detail. Suppose that a group needs to form a group opinion on three propositions by voting on them individually. The basis of their individual beliefs is a valid argument expressing a relation between the propositions. The argument contains two premises, p_1 and p_2, and the conclusion p_3. Since the argument is valid, any rational agent will recognise that, if p_1 and p_2 were true, p_3 must be true as well. This relation can be expressed by an implication relation: $(p_1 \wedge p_2) \Rightarrow p_3$. Now suppose that three agents A1, A2 and A3 have individually consistent profiles as in Table 1.1.

Table 1.1: 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

The individual beliefs are rational in the sense that they are compatible to the validity of the argument: they all accept that, were the premises true, so would be the conclusion. But the majority opinion aggregated through sentence-wise majority voting violates this constraint. The majority opinion continues to accept the validity of the argument, and also accepts all premises, but it denies the conclusion. The aggregated opinion is irrational in this sense.

Other examples for independence phenomena need not have the rationality of the group opinion as a condition of optimality. Instead, one could argue that a different collective optimum is given by a state of belief convergence. This is particularly interesting in human interaction involving learning, experimentation, and deliberation. Should rational and productively-minded agents engage in these activities for a sufficient amount of time, in the optimal case we would expect them to reach agreement sooner or later: problems in science find an accepted solution, and political discussions are brought to an agreement that the vast majority accepts. In these cases, the collective optimum would be described by near-universal agreement. It turns out, however, that individually rational agents can end up polarised through prolonged argument exchange. To be more specific, when agents in isolation successfully find supporting arguments for their beliefs, they attain consistent and well-confirmed beliefs. But when multiple agents strive to confirm their beliefs in this way, a group of such agents is not unlikely to enter a polarising dynamic – even if all agents are always responsive to arguments offered by others. In this thesis, Chapters 3 and 4 are about independence phenomena with the second kind of collective optimum, and Chapter 5 about the first.

The two examples just mentioned involve discrepancies between individual rationality and a failure to achieve collective optima. But independence phenomena can run the other way around, as Zollman (2010) shows. In these different examples, the collective optimum is in fact achieved, but through individually *irrational* behaviour. This time, it is the individual irrationality that is responsible for the mismatch between individual rationality and the collectively optimal state.

Examples for independence phenomena of this group can be found in the history of science and engineering. Consider that initially well-supported scientific theories occasionally turn out to be false as science progresses. Further suppose that to be individually rational, scientists should accept and continue to work on the theory that is best supported by their current evidence – call this the "evidentialist norm". Then, until new evidence surfaces, all scientists are individually rational in accepting the received theory, and any scientist who is pursuing an alternative might be taken to act in an irrational, even outlandish fashion. But in some cases this seemingly outlandish behaviour bears fruition and the improvement in scientific theory convinces not just the other scientists, but contributes to the well-being of all humankind. Examples can be found in the history of plate tectonics (Kitcher 1990, pp. 7–8), peptic ulcer disease (Zollman 2010, pp. 19–22), and the development of blue LEDs (Nakamura, Pearton & Fasol 2000, §2.6). We now know that the continents of our planet

rest on plates and that these plates shift, but the evidence available to earlier geologists made seem an alternative theory more probable. We now know that bacteria are a major cause for peptic ulcer disease, a disease in the human stomach. But earlier evidence erroneously suggested that bacteria could not survive in the human stomach. And while it seemed most promising in the 1990s to manufacture blue LEDs from zinc compounds, we now illuminate the world with blue LEDs manufactured from gallium compounds instead – even though that approach was described by researchers as "hopeless" and by its own inventor as a "gamble" (Nakamura, Pearton & Fasol 2000, p. 17). All of these cases involve major breakthroughs in science and engineering that were facilitated at least in part through individual deviation from the evidentalist norm.

1.3 The many faces of rationality

Rationality is a cover term for higher cognitive functions, including belief acquisition and revision, planning for goals and desires, or deliberation and reasoning (Knauff & Spohn 2021, p. 3). Rationality can be studied empirically, as in the biological foundations, activation and impairment of rational functions in humans and other agents. Call this the descriptive-empirical study of rationality. But rationality can also be studied normatively as in general rules for the optimal use of higher cognitive functions. Individuals are rational in this latter sense if they apply their higher cognitive functions in a way that is compatible to such general rules. For this normative study, of course, a main objective will be to figure out what exactly these general rules for the application of higher cognitive functions consist in.

It is common today to partition the normative study of rationality into the fields in which agents apply their cognitive functions. It is common to speak of *economic* rationality, *practical* rationality, or *epistemic* rationality. Epistemic rationality is the application of higher cognitive functions insofar as they are intended to yield knowledge, justified belief, and related goals. These fields of application may overlap or they may conflict. For example, what is rational for you to do economically might depend on what beliefs you should rationally hold. In this thesis, however, I am exclusively concerned with epistemic rationality, and I will use the terms "rational" and "rationality" to apply to this sense only.

Being rational is not always the same as being right. You can be epistemically rational without being aware that you are right, and you can be rational while being wrong (see Comesaña 2020 for a discussion of cases like this). The truth is not always decisive in deciding whether an agent is epistemically rational, and epistemic situations can permit multiple rational yet irreconcilable responses. These situations are called (epistemically) "permissive" (Schoenfield 2014, pp. 196–197). In this thesis, I will be exclusively concerned with epistemically permissive situations, the assumption being that humans are confronted with these situations on an all too regular basis.

With truth out of the equation, it might seem that permissiveness could entail that anything goes. The normative study of rationality thus needs to find criteria that do not rely on truth but still designate some responses to an epistemic situation as rational and others not. The literature offers two basic approaches to these general rules:

- *Structural epistemic rationality* (e.g., Kiesewetter & Worsnip 2023, §1.2):
  An agent a is epistemically rational in believing a proposition p if, and only if, that belief is coherent with a's other mental states.

- *Substantive epistemic rationality* (e.g., Lord 2013, §2.3):
  An agent a is epistemically rational in believing a proposition p if, and only if, in believing p, a is responsive to the available evidence.

These two characterisations are mere sketches of views that are well-developed in the literature. For example, Fogal & Worsnip (2021) discuss a variety of interpretations of the substantitive approach, and with a mind to application, show that even within substantitive theories, there can be deviation about which beliefs should be designated as rational.

The theoretical literature on epistemic rationality has both proponents and critics of the structural as well as the substantitive interpretation. For example, Broome (2021) maintains that an agent's rationality must only depend on this agent's mind – a criterion that is obviously met by structuralist but less so by substantivist approaches. On the substantive side, Heinzelmann (2022) argues that the structuralist approach can make conflicting demands that could not be simultaneously met by any agent. In these situations, structural accounts would be unable to flag any behaviour as rational or irrational. Substantive accounts, Heinzelmann says, can give definitive guidance in these cases. It would seem that substantive theories can handle more cases, giving these theories an advantage over structuralist ones.

As will become clear in later chapters, it is very important to the computational models presented in this thesis that the agents act epistemically rational. However, in this thesis I could not possibly offer any theoretical advancement on the theoretical question itself. The theoretical literature is concerned with understanding conflicting approaches to normative rationality – whether one is more fundamental than the other, and which one explains norms of rationality better. In this application-oriented work that I present here, the best way forward seems to accept, tentatively, both conceptions as describing distinct phenomena and verify agent behaviour for constraints of both types. The hope is that further theoretical developments will not invalidate the results, or require only minimal modifications. So in this thesis, I will verify that agents are structurally coherent as well as substantively responsive to the presented arguments.

With that being said, I do not wish to say that the distinction between structural and substantitive interpretations would not matter. The difference is not hard to make out. Imagine you would be working on a difficult problem in

your area of expertise. The problem may be so hard that it remained unsolved for years or even decades. Yet you make some progress on this important issue and thereby conclude that you want to tackle it. You expect that a solution or even a substantial advancement will bring economic rewards – a promotion, maybe. The next day you get a call for a job opportunity that would allow you to move away from your current dissatisfaction and economic disadvantage. After thinking about the offer, you decide to decline, since this rewarding yet also highly demanding job would prevent you from solving the important problem and thereby receive even more satisfaction and, presumably, wealth. It is not implausible to assume that you have reached a coherent state of mind: as far as your mental states are concerned, you have your eyes set on something important, you feel competent to tackle it, and taking the early off-ramp instead of continuing the journey is, you now believe, a foolish waste of opportunity.

But the substantive epistemologist is not satisfied. How reasonable is your belief that you will be able to make substantial progress on the important issue after years and decades of stagnation? It is not implausible to assume that the evidence available to you points to many failed attempts in the past and you have no specific roadmap to propose. You might have formed a coherent state of mind but that was only possible since you did not respond to essential pieces of evidence that would have been available to you. Your belief that the new job would distract you from your true calling is in fact irrational.

As I will verify in later chapters, the agents from the models described in this thesis adhere to structural and substantitive standards in forming their beliefs, and they revise their beliefs in accordance with rules deemed rational in the literature. They thus do exactly what the structuralist and substantive norms ask of them. But here lies a further theoretical issue of rationality. The question is whether structural and substantitive approaches cover all norms of rationality. The norms I have presented so far only govern how agents should form or revise beliefs given a fixed set of evidence. But they do not govern how agents should pursue the collection of new evidence. Call this process "inquiry" and philosophical questions about these processes "zetetic" (Friedman 2020, p. 501, fn. 1). In the example above, the norms do not specify how you should collect new evidence about the job or your important question. Unfortunately, it will remain an open question in this thesis whether agent behaviour meets the standards of recently discussed norms of inquiry, or "zetetic norms" (Friedman 2020). However, there is some reason to believe that this can be considered an independent question (Thorstad 2021, pp. 2919–2922).

1.4 Concepts of polarisation

In most general terms, polarisation is the formation of groups that become more cohesive internally while simultaneously diverging from other groups that are themselves increasingly cohesive. Polarisation in the political realm is a familiar issue to many people alive in 2024. In the United States, polling

data suggests that Republicans and Democrats are increasingly unlikely to socially interact, such as in marriage or friendship (Pew Research Center 2014, 2017). In Germany, polling data suggests wide gaps between followers of some, though not all, of the parties currently represented in parliament. The most pronounced animosity is attested for voters of *Alternative für Deutschland* and *Die Grünen*. In 2021, 77% of respondents who associated themselves with the first disliked the second, and 92% of those who associated themselves with the second disliked the first (Roose 2021, §7.3).

These cases involve the formation of clusters in humans that grow increasingly apart from those in other clusters. The existence of these clusters need not necessarily imply polarisation of the whole population. In the United States, many people have a low level of partisanship and value cross-party cooperation. In Germany, according to surveys at the time of writing this thesis, less than a third of the population considers voting for either of the parties mentioned above. Still, these pockets of polarisation can affect a population in its political and social procedures, such as through the demeanour that agents express in parliament, in the media, or on digital social networks.

It is sometimes assumed that rising polarisation was a characteristic feature of politics in the 21st century. But the available evidence does not seem to point in that direction. In the global context, the available data is mixed and does not indicate a global trend of polarisation. Some countries see polarisation dynamics comparable to the US, but others have recently experienced stagnation and some even a fall in polarisation (Boxell, Gentzkow & Shapiro 2024).

While it appears to be an urgent problem in some countries today, polarisation and clustering phenomena in humans are not at all unique to the 21st century. There is evidence that issues similar to what we call "polarisation" today troubled scholars in earlier times, particularly in the 18th century. David Hume's 1741 essay *Of parties in general* is a remarkable piece of evidence for earlier reflection on clustering phenomena in humans:

> When men are once inlisted on opposite sides, they contract an affection to the persons with whom they are united, and an animosity against their antagonists: And these passions they often transmit to their posterity.

In the essay, Hume reflects on disagreements of participants on "opposite sides" who form positive attitudes to their in-group peers and negative attitudes to others, somewhat similar to how we study cross-party affections today. In-group affection is not the only factor that Hume considers for clustering phenomena in humans. He also mentions that they might arise due to difference of opinion, such as different religious and political "principles":

> Where different principles beget a contrariety of conduct, which is the case with all different political principles, the matter may be more easily explained. A man, who esteems the true right of government to lie in one man, or one family, cannot easily agree

> with his fellow-citizen, who thinks that another man or family is possessed of this right. Each naturally wishes that right may take place, according to his own notions of it.

Note that Hume does not seem to be interested in disagreement in general, but in disagreements characterised by maximum opposition, fierce adversity, and those that "produce the greatest misery and devastation". Hume goes on to hypothesise that humans end up in these situations not as a matter of historical accident, but due to human nature itself:

> But such is the nature of the human mind, that it always lays hold on every mind that approaches it – and as it is wonderfully fortified by an unanimity of sentiments, so is it shocked and disturbed by any contrariety.

Hume offers some early reflection on polarisation and clustering phenomena, citing many potential examples from his contemporaries and from human history, and he also offers a typology of causes for these phenomena, including personal affection and difference of opinion. As I will show shortly, this typological approach to polarisation is shared with current approaches to polarisation in the social sciences.

There is more evidence for earlier reflection on polarisation and clustering phenomena in humans. At the time of drafting and dissemination of the constitution of the United States, the Founding Fathers published a series of 85 essays in newspapers that came to be known as the *Federalist Papers*. In the tenth instalment of this series, James Madison (1787) reflects upon the risk of faction in a democracy. Like Hume, Madison does not assign this risk to his contemporaries or compatriots alone, but identifies it as a risk in human nature overall: "the latent causes of faction are thus sown in the nature of man". Madison assigns more devastating consequences to factions than we would to polarising groups today, but he still reflects on clustering phenomena in humans and its different causes, including "different opinions concerning religion, concerning government" or "an attachment to different leaders ambitiously contending for pre-eminence and power".

This evidence for early reflection on clustering phenomena in humans suggests that academic interest in polarisation need not be motivated solely through the accidental situation in the early 21st century. Clustering phenomena are worthy of our study because they arise in many different contexts. A tendency to polarise is "sown in" (Madison) the very "nature of the human mind" (Hume). The study of polarisation thus offers a look into fundamental aspects of human reasoning and social interaction.

Early scholarship categorised clustering phenomena into different types. The types we use may differ today, but current sociology also follows a typological approach to polarisation. Recent contributions to sociology suggest that the term *polarisation* should not describe a single phenomenon, but that it is best understood as a cover term for a collection of concepts (Iyengar, Sood

& Lelkes 2012; Iyengar & Westwood 2015; Mason 2013, 2015). The literature describes three specific ways of growing apart: affective polarisation, polarisation of issue positions and group polarisation.

Affective polarisation is characterised by increasing animosity between groups each defined by a shared identity. It is sometimes called "social polarisation". An example for affective polarisation is the recent polling data from the United States mentioned above, suggesting that Republicans and Democrats become increasingly unlikely to socially interact in marriage or friendship (Pew Research Center 2014, 2017) – a clear indicator of affective polarisation in the United States. There are several ways to quantify affective polarisation. Boxell, Gentzkow & Shapiro (2024, §2) understand affective polarisation as the aggregated differences in respondents' affect towards 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 towards other individuals – but they do not poll their issue positions. For good reason: 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. We can also call it "belief polarisation". Bramson et al. (2017) collect different interpretations of this phenomenon: it could mean a rise of variance among the beliefs held in a population, but maybe the most comprehensive interpretation is how belief-based 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. This is the conjunction of features 1 and 2 in Esteban & Ray's polarisation concept (Esteban & Ray 1994, p. 824).

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 not only to politics. 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). Physicists can be drawn to opposite sides regarding the interpretation of quantum mechanics, and linguists can polarise about the question to which extent data from Amazonian languages refute universal theories of grammar. 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.).

Issue polarisation is about agents' stances towards on-topic claims, while affective polarisation concerns agents' attitudes towards 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 held (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 in this thesis.

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 this thesis, track agents' issue positions.

1.5 Concepts of diversity

Measures of issue polarisation are often aggregates of pairwise differences between beliefs, either in the sample-wide variation and deviation or as part of a cluster analysis. There is an approach to opinion diversity based on the pairwise differences between individuals as well (Weitzman 1992). However, this approach is now believed to face technical issues (van Hees 2004) and limited applicability (Nehring & Puppe 2002, p. 1158).

Instead, diversity is now most commonly measured in terms of how frequently types are expressed in a population (for a review, see Page 2011, Chapter 2). In ecology, the diversity of a sample can be understood as the frequency of species in this sample, where the individual animal's species serves as the type needed for the study of diversity. But diversity can be measured regarding many different traits of individuals, not just their species. A group that is described as "diverse" can have diverse demographic and economic backgrounds, diverse work experiences and competences, diverse problem-solving approaches or diverse viewpoints. All of these categories can serve as types the frequencies of which are needed to figure out a sample's diversity. In this dissertation I focus exclusively on opinion diversity, that is, the diversity of beliefs upheld in a group. And I understand opinion diversity in direct analogy to the measures of diversity in other disciplines.

What does it mean for a group to have opinions of diverse types? This clearly presupposes that the measured group can be partitioned into types based on their opinions – just like establishing ecosystem diversity depends on knowledge about the individuals' species. In Chapter 5, I will understand a type to be a cluster of opinions. The diversity measures in this thesis thus depend on an antecedent opinion clustering, as described later in Section 2.3.2.

Diversity can be understood as type frequency, but the frequency of types can be measured in two fundamentally different ways. It can be measured by either counting the number of types (absolute frequency) or by determining how many individuals express each type relative to the complete number of individuals (relative frequency). These two ways of counting give rise to two different diversity concepts: absolute frequency yields *richness* and relative frequency yields *heterogeneity* (Nehring & Puppe 2009).

To illustrate how heterogeneity approaches to diversity work, suppose that you mingle with the participants of an ethics conference. If you were to understand the crowd's heterogeneity in terms of the Gini–Simpson diversity index, you would inquire about the probability that you would 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. In this thesis, I will refer to "diversity" as the relative type frequency indicated by the Gini–Simpson diversity index, or the probability of encountering an individual that expresses a different type.

Both richness and heterogeneity are concepts of diversity, but they represent different phenomena and are not equally useful in all applications. In particular, they are not equally meaningful for smaller sample sizes. Richness is known to be unstable and hard to interpret in small sample sizes (Tuomisto 2010, p. 854). This is why I will later argue that diversity in the sense of richness would yield misleading results in the context of dialectical structure models. There are two reasons why simply counting the number of types in small samples can be misleading. Consider first a sample with 15 individuals, 7 of which express type A and 8 others express type B. The richness of this sample is 2 as two types are expressed in the sample. If we were to use the Gini–Simpson index to measure heterogeneity, we would obtain a diversity indication of 0.498. Now consider that we add two more individuals to the sample, one of type C, and one of type D. The richness of our sample will have doubled to 4, but the heterogeneity as expressed by the Gini–Simpson index would equal 0.60, or a rise in only 0.1. There is nothing wrong with saying that the richness of the sample has now doubled, but it is easily seen to over-estimate what has actually happened.

A second, potentially problematic consequence of absolute frequency indicators of diversity is that addition of further individuals can never make samples less diverse (Nehring & Puppe 2009, pp. 313–314). Any additional individual will either represent an existing type or add the presence of a new type. But samples can become less diverse, or "more homogeneous", in the sense of heterogeneity when new individuals are added. When new individuals express a type already present in the sample, their addition increases the relative share of this type, lowering the frequency of the other types and thus the heterogeneity of the sample as a whole. Whether this constitutes a problem will of course depend on the specific question that we tackle using diversity indices.

Diversity can be understood to mean different things. The agent-based dialectical structure models studied in this thesis have a small sample size of less than 100 individuals. For this sample size, the literature recommends the Gini–Simpson index as the most reliable (Tuomisto 2010, p. 854). In this thesis, I will

understand opinion diversity exclusively in this sense. It indicates the probability that the opinions of two randomly drawn agents belong to two different opinion clusters.

1.6 Why pursue social epistemology through argumentation?

So far I have introduced independence phenomena and concepts of belief polarisation and opinion diversity. The simulation experiments reported in this thesis study independence phenomena under different expressions of diversity and polarisation. In these simulation experiments, agents 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. Computational models of argumentation yield substantial results that can improve our understanding of independence phenomena, and argumentation seems to have explanatory power in matters of social epistemology. But that is not the only reason why models of argumentation are worthy of pursuit in social epistemology. There are antecedent, systematic considerations that explain how argumentation fits into the study of the social dimensions of knowledge and justified belief.

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 & 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. And Cartwright's theory also captures disagreements about 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 towards the guilt of the defendant (e.g., because such a knife is widely available).

Humans, of course, usually deviate from providing arguments in formulaic language, like $(p_1 \wedge \ ... \wedge \ p_n) \Rightarrow c$. 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. It is not necessary for my present purpose that arguments are present verbatim in all instances of sharing evidence. I only maintain that they serve as an adequate *abstract representation* of this process.

To summarise, both Mercier & Sperber (2011) and Cartwright (2013) advance substantial theories about what reasoning and evidence fundamentally are, although nothing in the following requires accepting strong readings of these theories and their ontological commitments. The benefits of studying argumentation in social epistemology can be recognised without buying into these commitments. What I take these theories to indicate is that we can consider argumentation as a useful *representation* of deliberation, reasoning and evidence exchange.

Concerns about this picture might emerge from the fact that arguments often fail to change minds. If arguments and argumentation had epistemic import, shouldn't there be a more steady and linear effect of argumentation on agents' beliefs? Contrary to the optimistic outlook that philosophers at times assign to arguments, it is not the case that they always pave a smooth path to consensus (for a discussion, see Dutilh Novaes 2023). Should this make the prospect of studying argumentation for epistemology less attractive? It might seem so: if argumentation is not a consensus machine and so often fails to bring about warranted change of mind, then maybe we should focus our attention on the processes that influence belief dynamics more reliably?

A first response to this worry could be that we restrict ourselves to situations in which agents accept the force of the better argument. In fact, the computational models presented below assume that artificial agents always accept the validity of all presented arguments. In many cases, though not always, this leads agents to change their beliefs. Just like Bayesian epistemologists assume that epistemically rational agents always conditionalise on their priors, models involving argumentation assume that rational agents change their views whenever suitable arguments are presented.

But the theories of evidence and reasoning reviewed above suggest a second, different response. According to them, core epistemic processes and phenom-

ena are argumentative in nature and they can be studied through studying argumentation. This is a characterisation of the nature and operation of these processes, but whether arguments change minds is a separate question from our fundamental look at these processes.

Argumentation is not just a tolerated, but an *expected* behaviour of humans in many social interactions and institutions. Human interaction in science, court, or parliament is fundamentally argumentative. But not every interaction in these institutions is followed by belief change. For example, a research paper that does not convince you still contains many arguments through which the position and reasoning of its authors can be understood. From this point of view, those who consider belief change to be the primary purpose of argumentation misjudge its function.

We should avoid a nihilistic outlook on argumentation when it does not bring about belief change. Instead, engaging with the arguments offered by others offers an indispensable outlook into their beliefs and reasoning. Suppose that you learn that someone beliefs q, a proposition you find implausible, and that this is all that you know. Now consider the slightly modified scenario in which you do not only learn about the belief in q, but also the argument that q because of premises $p_1, p_2, ..., p_n$. The scenario in which you learn not just about the belief in q but also about the other beliefs and their inferential connection to q carries more potential for understanding and interaction. Maybe the ps even express beliefs that you share? In humans, the core mental faculty of argumentation offers an outlook into their reasoning – irrespective of whether any particular line of reasoning convinces others.

A second way of being nihilistic about argumentation is to think that others who are opposed to one's own view "had no arguments", or were incapable of arguing rationally. This view is implausible in light of Mercier & Sperber's theory about argumentation and human reasoning. If argumentation is a core cognitive function in humans, we should be ready to acknowledge the validity and sincerity of arguments we observe in most circumstances. After all, it would seem odd to deny the adequate performance of other cognitive functions, such as uttering grammatically correct sentences or finding the correct answer to arithmetic problems.

Whether one arrives at an optimistic or a pessimistic outlook on belief change through argumentation might also depend on the scenarios that one envisions. When two experts that are already acquainted with many arguments meet, a further argument announced by either expert does not necessarily guarantee a change of mind. People with a high degree of partisanship of opposing persuasions easily find themselves in the same situation. Compare this to a teacher who is conveying new information to students through argumentation, such as presenting a geometric argument for Pythagoras' theorem. A change of mind seems more natural and plausible in the latter case than it does in the former. We should neither adopt an overly optimistic nor pessimistic outlook on argumentation, but acknowledge its role in the formation and dynamics of belief in ourselves and our interlocutors.

It is not at all a modern nor a uniquely Western idea to consider the role of argumentation in the rational acquisition of knowledge and justified belief. There is evidence that some philosophers in ancient India considered this connection (Lorenz 2008, pp. 94–96). In ancient Greece, the connection is particularly clear and developed in the work of Aristotle (Barnes 1969; Striker 2022). Today, it offers us a promising outlook into epistemic processes and particularly their social dynamics.

1.7 Opportunities and challenges in computational social epistemology

So far I have motivated the idea that processes of argumentation can improve our understanding of social epistemology in general and independence phenomena in particular. This thesis is about the study of independence phenomena in argumentation-based computational models. In this section, I motivate and reflect upon the use of computational modelling as a method in social epistemology. After considering methodological arguments about computational modelling received from the literature, I present a different kind of argument according to which computation is useful since it allows us to study dynamic aspects of philosophical concepts and normative constraints. A consequence of this argument is that computational approaches are much less a disruption and much more of a continuation of philosophical methods. In the second half of this section, I review challenges that arise in the interpretation of results from computational models.

What is a computer simulation, and what is a computational model? Computational models are representations of a target system in a computer, together with its environment, states, processes, and agents. Simulations on computational models progress through what Winsberg (2022, §1.1) calls "rules of evolution". These rules are implemented in the computer to transform the representation of the target system from one state to a following state. Call this a "step" in a simulation. Some rules of evolution can be described through differential equations such as in numerical weather prediction. But not every rule of evolution can be described as a differential equation, particularly in agent-based models. In agent-based models, the rules of evolution are formal descriptions of agent behaviour, both towards their environment and other agents. A simple rule could be, "consider all opinions in the group and update your own by meeting halfway the agent with whom your opinion overlaps most".

With epistemological goals in minds, agent-based computational models are particular useful to study the dynamics of belief systems. Agents are thus implemented by way of their belief system and the mechanisms with which they explore their epistemic environment and influence the beliefs of others. The factors that influence agents' beliefs can be divided into factors of "informational" and "social" influence (Burnstein & Vinokur 1977), where informational influence consists in signals that agents receive from the simulated world, such as results from an experiment, and social influence signals are transmissions

from the beliefs of others. Agent-based models aggregate the mathematically described beliefs and belief dynamics to properties of the whole artificial society. In social epistemology, these aggregations can result in measures of agreement, average closeness to truth, belief polarisation, opinion diversity, and other epistemically interesting properties. The study of these models in simulation experiments can then be an investigation into the evolution of these properties, and on the factors and conditions that drive their dynamics.

The beliefs of artificial agents and their behaviour are representations of cognitive functions and rational capabilities in a computer. We can use existing theories about rationality to verify that these representations are informative about our understanding of rationality and belief dynamics in humans. But outside of the boundaries described by our concepts and theories, the behaviour of artificial agents can be very much unlike human behaviour. Artificial agents do not show variation in character (unless we model them in this way) and are not subject to many other aspects of social life in humans. If artificial behaviour is reduced to functions of rationality, can we learn anything at all from these computational models – particularly, without comprehensive empirical validation? My answer in this section will be that we can. Simulations on artificial agent behaviour provides new evidence about the dynamic aspects of formal concepts and normative constraints.

Computational modelling is an emerging if still unusual method in social epistemology and philosophy as a whole. Why should philosophers accept the pursuit of computational models in their discipline? In the following, I discuss four arguments about the systematic exploration of computational models in philosophy. These arguments are necessary because of an "often implicit" sentiment that computation was "not philosophical" (Mayo-Wilson & Zollman 2021, p. 3649). This sentiment should surprise us for two reasons. First, it does not seem to follow from diligent methodological consideration. If logic and probability theory are admissible tools in philosophy, which reasons are there to find computation inadmissible? Second, this sentiment is likely to prevent methodological improvement in our discipline. We could do without continuous methodological reflection and improvement in a world with fixed philosophical methods and a fixed role for philosophy in relation to other disciplines and in society as a whole. But we do not inhabit such a world.

Hintikka (1988, p. 272) draws this connection between the methods that philosophers use and their ability to interact with other disciplines and society as a whole. Philosophy has lost relevance in the dialogue with other disciplines, or so he claims, because the knowledge that could possibly be gained from its established methods and practices has run out:

> Part of the bad news is that skills of general argumentation, including the articulation of our so-called "intuitions", are generally relied on by professional philosophers also when they tackle major research problems in their own discipline. This is bad news because skills in persuasion and argumentation simply are not

> enough to shed new light on the kinds of crucial theoretical problems which philosophers have been able to master in the past. As a result, what is known as philosophical research has lost much of its significance and its interest for scientists and scholars outside philosophy itself, with the exception of linguists. The contrast with earlier decades of this century is in fact quite striking.

Call this the argument from methodological exhaustion. While Hintikka admits that general argumentation and exploration of intuitive responses might have been productive in the past, this well has now dried up and no further insights can be gathered from these established methods. This is not a direct argument for the use of computational methods, but it motivates a culture of appreciation towards any new philosophical method.

The closeness of philosophy to science provides a second indirect argument in favour of computational modelling, a view that Thagard (2018, p. 462) explicitly mentions. For those that view philosophy as continuous with science, it might seem plausible that this continuation can be established, in part, through the adoption of scientific methods. Introduction of computational models to philosophy should thus seem permissible since computer simulation is an accepted method in many fields of science. But of course there are many different scientific methods – think DNA sequencing, radiocarbon dating – and not every scientific method has the same utility in every field. Like the first argument, it is an argument that opens up the possibility but does not establish conclusively if and why computational modelling is useful in philosophy.

The third argument in favour of computational methods in philosophy is of comparative methodological nature. It proceeds by considering an established method and then by arguing that simulations on computational models can serve philosophical interests at least as well. Thagard (2018, p. 463) makes a particular forceful presentation of an argument of this type:

> They [computer models] are thus much more useful than thought experiments, in which philosophers' own intuitive reactions to stories they have made up are mysteriously used as evidence for the philosophers' preconceptions.

A comparison between thought experiments and computational modelling is interesting because both methods produce hypothetical scenarios. These scenarios are evaluated against the intuitions of the reader in thought experiments, and through computational analysis and measurement in modelling. But I am not convinced by the comparative argument as Thagard (2018) presents it. It is true that philosophers' preconceptions are the first thing that philosophers should question, and it is also true that implausible thought experiments help no one. But the worst instances of a method are not indicative of a method's general usefulness. It would be undesirable to abandon insightful and pedagogically valuable thought experiments such as Hilbert's Hotel or Rawl's Original Position just because there are also bad thought experiments.

There are more reasons why Thagard's very general disapproval does not convince me. First, the prejudices encoded in thought experiments might just as well be encoded in computational models. They are more accessible to reflection and criticism if they are put into formalisms and publicly accessible computer code, at least in theory, but the nature of prejudice is that it hides in plain sight. Second, looking around us, we will see that many scientific disciplines today accept a diversity of methods in their progress. There is no apparent reason why philosophy should not accept a diversity of methods as well. Computer modelling might sometimes be helpful and sometimes not, just as thought experiments might serve some purposes well while being useless in others. A case-by-case analysis seems much more fruitful than a general ranking of methods.

Mayo-Wilson & Zollman (2021) put forward a different version of the comparative methodological argument. They argue that philosophical thought experiments serve at least six different purposes (Mayo-Wilson & Zollman 2021, Section 1.1), and then find that computational modelling in philosophy can serve five of these at least as well. The remaining goal they find thought experiments to handle better is the elicitation of normative intuitions in hypothetical scenarios (think trolley cases). Computational modelling, they argue, is better suited to understand the logical relations between theses or explore social dynamics, even though thought experiments are also sometimes used for this purpose. These findings of Mayo-Wilson & Zollman (2021) are valuable as they show the potential benefit that philosophers can hope to achieve through computational modelling. But whether a particular computational model meets any of these goals will depend on the specific case. There might still be cases in which thought experiments surpass computational models of the same target phenomenon.

To Mayo-Wilson & Zollman's argument I would like to add another argument for the use of computational modelling in philosophy. According to this argument, computation is not in opposition to, but in fact the continuation of established methods in philosophy. Many of our established methods strive for the clarification of norms, demands, and concepts, and do so often in mathematical or otherwise formal ways. These are ways that are naturally extended to computation, and with computation, the full nature and ramifications of these norms and concepts can be understood. This particularly applies to aspects and properties that can not be derived analytically from definitions and formalisms. For example, consider how Bayesian epistemologists explain how beliefs can be described through real numbers and with methods of the probability calculus (e.g., Titelbaum 2022, p. 12) and how rational requirements for belief updating can be formulated in the same framework (Titelbaum 2022, §4.1, §5.5). Since these concepts and norms refer only to terms that can be described in logical and mathematical formalisms, they can be studied computationally. For example, we can ask whether it is possible that agents would ever polarise in their beliefs as they follow these norms (the answer seems to be "yes", Dorst 2023). The discovery of epistemically rational belief polarisation in Bayesian agents improves our understanding not only of Bayesian episte-

mology. Insofar as Bayesian epistemology can be a characterisation of human epistemic behaviour and rationality in general, this discovery also improves our understanding of epistemic dynamics in humans and the abstract nature of rationality. As Williamson (2017) says, insofar as our formalisations of rationality are adequate, discoveries about our formalisms are also discoveries about rationality in general. Note that empirical verification is less important for this particular use of computational modelling compared to verification against our concepts and constraints.

In short, this argument relies on the ubiquitous use of formalisation in philosophy to emphasise that their computational study provides opportunities to understand their dynamic aspects. These dynamic aspects of the formalisations and, by extension, the target phenomena they represent, would not be accessible by analysis alone. When we allow ourselves to view computational modelling from this perspective, it is much more a continuation than a disruption to philosophy. We can now regard any formalisation of a philosophical problem as a computational model waiting to be implemented.

Data can be used for multiple purposes. It can improve our understanding of the world around us, but it can also be used to guide and evaluate our actions. Consider this important example. In 1854, physician John Snow visualised cholera cases on a map of the West End of London. The data indicated that the local outbreak clustered around a particular public water pump. This finding improved our understanding of the world around us, as it contributed to establishing that cholera is transmitted through contaminated water. But this data also guides our actions – civil engineering should separate freshwater supply and sewage. The computational data presented in this dissertation is of less immediate and dramatic nature, but it also serves these two purposes. While my priority is to show how computational data improves our understanding of epistemic rationality, argumentation, and uncertainty, the same data can also be used to guide and evaluate our behaviour. I will suggest some normative guidance on the basis of this new understanding, but I will do so mindful of the fact that the underlying data could be interpreted differently.

As of writing this dissertation, the constraints and rules that govern the application of computational modelling in philosophy are still to be determined. But it is obvious that the benefits of modelling will not be gathered from any arbitrary application of computational methods. At the very least, computational methods need to be supported through a faithful implementation of the formal apparatus as well as the target phenomenon, and there needs to be sufficient robustness analysis under all plausible parameters of a model.

But when such standards become available, meeting their requirements will not be the only challenge to the use of computational modelling in philosophy. A major source of uncertainty is the reliability of the conclusions drawn from computational data, a debate that is not new to disciplines like sociology, too. Some challenge the view that computational modelling will provide explanations of real world phenomena because computational methods can not trace back the causes that *actually* led to an event and that could predict future instances (Grüne-Yanoff 2009). Although Grüne-Yanoff does not discuss models

in social epistemology, it is easy to direct his critique to these models, too. Assume that it was possible, for a given group of humans and a given discussion, to obtain opinion dynamics in an agent-based model that was a good fit of their actual dynamics. Could we then conclude that the agent-based model uncovered a causal explanation for the observed dynamics? No, says Grüne-Yanoff. And there is some merit to this critique seeing that even empirically oriented applications of computational modelling might be better suited to explain social dynamics after the fact compared to predicting its occurrence. For example, consider how Friedkin et al. (2016) were able to reproduce opinion dynamics of the U.S. public regarding the 2003 invasion of Iraq long after the fact, but not before public opinion changed so dramatically.

Elsenbroich (2012) defends computational modelling of social phenomena against Grüne-Yanoff's critique. Grüne-Yanoff makes, according to Elsenbroich, unrealistic demands about explanations of social systems. Theories about these systems are not predictive in general, irrespective of whether they are gathered empirically or computationally. Elsenbroich's argument has some merit, too. For example, even though empirical social science has taught us some factors that determine a voter's choice in a parliamentary election, we are still unable to definitively predict the outcome of next week's election. To expect predictive power of computational models is to raise a demand that none of the other approaches to the study of social systems meet.

A different legitimisation of computational modelling in social systems could be that these models do not yield "how actually" explanations of a phenomenon, but "how possibly" explanations (Frey & Šešelja 2018). These explanations would not contain deterministic predictions of events, but rather describe potentially occurring events and counter-factual alternatives that could have materialised instead of the actual chain of events. Rosenstock, Bruner & O'Connor (2017, p. 239) go a bit further and even deny, at least for some models, that they provide how possibly explanations. They see the contribution in "stories" (not explanations) of how a phenomenon occurs "potentially". According to this approach, some agent-based models do not explain events in the real world at all, they only allow us to generate first hypotheses for further exploration.

This discussion about predictions, possibilities, and potentialities is about events that actually occur, how they could possibly occur, or what their potential causes are. This discussion is very much tied to the idea that computational models should be about events in the real world. But that is not the only way in which modelling can improve our understanding. Above I said that computational models implement formal representations of philosophical concepts and constraints and can provide new insights into their dynamic aspects. Through these dynamic aspects, we can improve our understanding of the demands these formal constraints make of agents. The discovery of belief polarisation under condition of epistemic rationality can be considered a conceptual truth about our concepts of rationality and polarisation. A computational model is always a model of something, but it does not need to be a model of actual events.

I take the models of belief polarisation from Chapters 3 and 4 to be of this conceptual explanatory power. These models show that plausible assumptions about epistemic rationality do not eliminate the possibility of belief polarisation. While this might be relevant for some events in the real world, the explanatory benefit of this model is of conceptual value. We learn something about the concept of epistemic rationality from this model – and, by extension, about higher cognitive functions in humans.

The interpretation of the model in Chapter 5 is more difficult. The target phenomena are events of majority voting, and simulations on the model reveal that highly diverse groups have a significantly reduced chance of maintaining consistency in this particular aggregation procedure. It is noteworthy that this model represents majority voting not through an intermediary process – voting is the same for artificial agents as it is for humans. But I will not argue that we should expect diverse groups of humans to always fail in making decisions when a sentence-wise majority vote does not succeed. Humans, after all, have infinitely more flexibility than artificial agents in responding to uncertainty. I will use this theoretical possibility instead to motivate reflections about the expectations we hold towards diverse groups that use this aggregation procedure as part of their decision-making. This improves our understanding of exceptional situations involving uncertainty and a pressure of time.

On a final note, the question about realism in computational models is particularly interesting regarding the agents. The question is simple: who are they? Whom do they represent? Agents in computational models of social epistemic processes are reduced to their rational capabilities. These entities are governed solely by the regularities and prescriptions of the formal model, and they are expressions of different behaviour allowed by the model. In the beginning 21st century, it does not seem far-fetched to believe that these functions of rationality could be realised in artificial systems, like content-generating machine learning applications or in automated decision-making. This dissertation can be read to be about epistemic dynamics in humans, but it might just as well be read as being about the dynamics in other rational agents. Humans sometimes are quite like the agents in this thesis, but sometimes humans are also quite unlike them. Even if we can derive knowledge about human rational behaviour through the study of computational models, we should expect these effects to co-occur with other influences, such as personal affections and emotions, preferences and economic standing, social roles and duties, etc. When the agents in computational, agent-based models tell a story, they do tell a story about how humans sometimes are – but they do not tell a story about how humans always are.

Having now motivated a computational approach to study independence phenomena in social epistemology, and after advertising its opportunities while reflecting on the challenges it comes with, I now move on to present the actual computational framework used in this thesis (Chapter 2). Two applications of this framework are later presented in this thesis: to study phenomena of belief polarisation, in Chapters 3 and 4, and to study opinion diversity and belief aggregation in Chapter 5.

Chapter 2

Modelling debates in the theory of dialectical structures

2.1 Motivation and origins

The interest in argumentation and its role in human interaction is shared cross-culturally and across many epochs of human scholarship from which written records survive. Ancient Chinese scholars distinguished argumentation from persuasion (Indraccolo 2021). In ancient Greece, parts of Aristotle's work are dedicated to the study of arguments and argumentation. The *Rhetoric* contains a distinction between the rhetorical use of arguments (persuasive speech) and its dialectical use (the inquiry into the truth of a statement). In the *Topics*, Aristotle provided what can be considered an early typology of arguments, distinguishing many different patterns and uses of argumentation.

A powerful tool available to the contemporary study of argumentation is the *standard form* of arguments (e.g., Feldman 2014, pp. 69–71). A standard form is the result of argument reconstruction, a process through which premises and a conclusion are extracted from arguments occurring in natural language. Standard forms are helpful to determine many features of an argument. For example, they can help determine whether an argument is valid. An argument is called valid just in case its premises lend sufficient support to the truth of its conclusion. The premises of an argument lend sufficient support to the conclusion just in case it is not possible for the premises to be true and the conclusion false. The standard form representation of arguments can be helpful in determining the validity of an argument, and to see whether implicitly assumed premises need to be spelled out for the argument to attain validity.

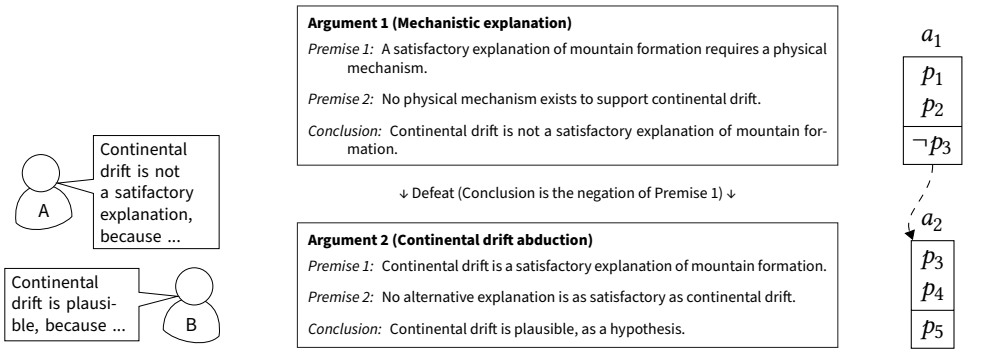
Standard forms are also useful to determine whether an argument instantiates a common pattern of argumentation. Arguments come in different shapes and sizes, and they are used in infinitely many different contexts. Some arguments justify the moral permissibility of an act. Other arguments justify the plausibility of a scientific theory. A fundamental discovery has been that argumentation in humans follows distinctive patterns, or "schemes" (Walton, Reed & Macagno 2008). Some share the common pattern of inferences to the best explanation, others the common pattern of reductio ad absurdum. When considering arguments through their standard forms, these common patterns can be read off from the reconstructed forms.

The standard form is helpful for the study of some aspects of argumentation, but the standard form does not cover all the aspects of argumentation and its social dynamic in humans. For example, it abstracts away from matters of style, the social roles of those who produce and those who evaluate arguments, and of the wider context that the arguments are placed in (Vorobej 2006, §1.1, §1.4). The standard form is instead geared to the optimal representation of the inferential, logical and semantic relations that hold between propositions that participants discuss (Betz 2005, p. 55).

The standard form is a tool for the analysis of single arguments, but it alone does not provide for understanding the relation between multiple arguments. Here is where the theory of dialectical structures steps in, the theory that forms the framework underlying of this thesis. The theory takes off from the observation that arguments usually do not appear in isolation. They are instead related to other arguments by what we can call *dialectical* relations (Betz 2009, p. 284). The theory of dialectical structures is not revisionary regarding the standard forms – it assumes that arguments are reconstructed in standard form and supplies relations and other modes of analysis for sets of reconstructed arguments. The theory of dialectical structures is thus best considered an extension to the study of arguments and argumentation.

The theory allows for the study of argumentation in agent-based models by abstracting from standard forms composed of sentences in natural language to symbolic sentence variables that can be handled in a computer. It is the goal of this chapter to introduce this way of modelling.

One can look at human argumentation in natural language, reconstruction of single arguments, and agent-based modelling in the theory of dialectical structures by way of increasing abstraction. Figure 2.1 illustrates these three levels of abstraction.



(a) Human deliberation (b) Argument reconstruction (c) ABM

Figure 2.1: Levels of abstraction

Two arguments and their standard forms are shown in panels (a) and (b) of the figure. Panel (b) also shows the two reconstructed arguments equipped with the defeat relation from the theory of dialectical structures. The theory explains why a defeat relation holds between the two arguments: the conclusion in the standard form of Argument 1 is the negation of a premise in the

standard form of Argument 2. I will explain the defeat relations later in this chapter, as well as other aspects of the theory of dialectical structures.

Panel (c) in Figure 2.1 shows the further symbolic abstraction needed for the implementation of an agent-based model in a computer. The premises and conclusions from the standard forms have now been replaced by sentence variables that can be modelled computationally.

Figure 2.1 also illustrates what we can hope to gain from agent-based modelling. In so far as the representation in (c) is a useful of abstraction of (b), and the reconstruction in terms of standard forms (b) is a useful abstraction of human deliberation in (a), the representation in (c) is also a useful abstraction of (a). By learning something about agent-based models (c), we may also hope to learn more about deliberation in humans (a).

The theory of dialectical structures is not the only framework that can represent relations between multiple arguments, and it is not the only theory that can be used for the implementation of arguments in agent-based models. The theory of dialectical structures is related to theories about *argumentation frameworks*. In fact, dialectical structures *are* argumentation frameworks of a particular kind. But let's take it step by step.

In logic and artificial intelligence, an argumentation framework is a tuple $\langle A, D\rangle$, where A is a set of arguments and D a defeat relation between arguments, $D \subseteq A \times A$ (Dung 1995, p. 326). Argumentation frameworks can be interpreted as graphs composed of nodes in A and edges D, as illustrated in Figure 2.2.

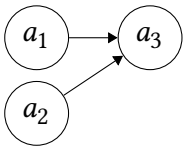


Figure 2.2: Illustration of an argumentation framework with three arguments. a_1 and a_2 defeat the argument a_3, as indicated by directed edges. In tuple notation, this abstract argumentation framework is described by $\langle\{a_1, a_2, a_3\}, \{(a_1, a_3), (a_2, a_3)\}\rangle$.

Considering that arguments may be defeated by other arguments, the conditions under which one may be permitted to maintain an argument is a main concern for research on argumentation frameworks (e.g., Besnard & Hunter 2008, §2.2). Argumentation theorists would say that an argument then is "acceptable". This interest is not pursued in this thesis.

Argumentation studied in terms of these argumentation frameworks is *abstract* in the sense that it does not account for the micro-structure of premises and conclusions in individual arguments. As I said above, arguments as they are studied in standard forms are complex objects that are composed of premises and conclusions. Abstract argumentation frameworks abstract away from this inner structure. But since they do not model arguments as complex objects, it is not always clear whether they are specifically models of argumentation

and not more general models of beliefs, reasons, or hypotheses. The theory of dialectical structures instead studies *structured* argumentation frameworks in which the micro-structure of arguments is always reconstructed in terms of premises and conclusions. Abstract and structured argumentation frameworks thus capture two different meanings of "argumentation". In the abstract sense, "argumentation" can mean the exchange of reasons, some of which defeat each other. Structured frameworks capture this sense as well, but they also give a very specific meaning to the kinds of reasons that are exchanged: arguments, understood as complex objects, represent an inferential activity from some propositions to others. They understand argumentation not just as the exchange of reasons, but the exchange of inferential information.

Abstract representations of argumentation are limited in a second way. While they capture the fact that arguments can be related to other arguments, many abstract argumentation frameworks only account for defeat relations among arguments (although diverse variations of defeat are known, Arieli et al. 2021, p. 1800). But arguments can also be used to *support* the reasoning in other arguments, and this support relation is independent from and irreducible to the defeat relation (Cayrol & Lagasquie-Schiex 2005). This support relation between arguments is indeed attested in the empirical study of argumentation in humans (Polberg & Hunter 2018). By convention, argumentation frameworks with one relation are called *unipolar*, while argumentation frameworks that allow both defeat and support are known as *bipolar*. Dialectical structures are structured bipolar argumentation frameworks.

The study of dialectical structures in agent-based modelling is thus motivated for two main reasons. First, because it continues the established practice of studying arguments as complex objects in terms of their standard forms, and brings this study to computational modelling. And secondly, because it allows for the study of bipolar relations in argumentation frameworks that seem indispensable for both systematic reasons (Cayrol & Lagasquie-Schiex 2005) and because the empirical study of deliberation in humans seems to make bipolar approaches necessary (Polberg & Hunter 2018).

The remainder of this chapter explores the features and mechanisms of dialectical structure models. Section 2.2 describes the main objects of these models, arguments, debates, as well as agents and how they form belief systems in light of the presented arguments. Section 2.3 describes measures that can be applied to these models, particularly those of agreement and disagreement, belief polarisation, opinion diversity, and of debate progress and normalised information entropy. Section 2.4 describes the dynamics of dialectical structure models. It describes how a debate progresses by adding more arguments to a dialectical structure, and how agents revise their beliefs. It also introduces the two main sub-types of dialectical structure models, one in which a debate evolves through iterative argument introduction (types 1a and 1b) and one in which argument maps are synthesised (type 2). Section 2.5 compares the dialectical structures approach to other agent-based modelling approaches.

2.2 Arguments, debates and belief systems

2.2.1 Arguments and validity

The theory of dialectical structures describes arguments as implication relations holding between a set of premises and a conclusion. This means that arguments are complex objects: they consist of a set of premises and a conclusion, and the premises $p_1, ..., p_n$ jointly imply the conclusion c, or $(p_1 \wedge ... \wedge p_n) \Rightarrow c$.

The premises and conclusions are drawn from a sentence pool of limited maximal extension. It contains propositional variables as well as their negation. Sentences that are premises in one argument can be a conclusion in another argument, and vice versa. The sentence pool is fixed in some variants of the model, but it can also be configured to be extended in the course of a model run. Even for expanding sentence pools a limit of maximum expansion has to be set.

In models presented in this thesis, all agents recognise all arguments in a debate to be valid. And when it is their turn to introduce new arguments, they only introduce valid ones. For any agent, accepting the validity of an argument means accepting that the truth of the premises is sufficient for the truth of the conclusion, or that it is impossible that the conclusion is false while the premises are true. The agents essentially maintain that the inference from the premises to the conclusion in fact holds. The point of disagreement in dialectical structure models is not whether the arguments are valid, but how the propositions in those arguments are to be evaluated. Agents may, and usually will, disagree on the truth of the premises and conclusions.

The term "validity" may appear suspicious in this context. After all, isn't validity a concept of strictly deductive arguments only, that is, arguments that hold purely due to logical relations (Dutilh Novaes 2021, §2.1; Feldman 2014, p. 75)? This sentiment that deductive arguments are only about abstract matters of logic, semantics, and mathematics can be called validity *in the strict sense*. Many arguments that humans make and that draw on other resources would not be valid in this strict sense. These arguments include inductive arguments, reasoning by analogy, or inferences to the best explanation.

Instead of this strict interpretation of validity, I would like to apply validity *in the broader sense* to understand argument analysis in the standard form and in the theory of dialectical structures. This broader sense allows validity for inductive and other non-deductive arguments. A long tradition has in fact regarded inductive arguments to be special cases of deductive argumentation. For example, Mill (1843, p. 514) thought that "the instrument of deduction alone is adequate to unravel the complexities" of induction. Russell (1903, p. 11) similarly called induction a "disguised" form of deduction. For further discussion of this idea see, e.g., Black (1966) or Graves (1974).

How can this tradition of understanding validity in a broader sense justify the acceptance of induction and other non-deductive arguments as valid ones? One strategy is to regard non-deductive arguments as enthymematic (e.g., Vorobej 2006, pp. 14–16). Enthymemes are arguments in which at least

one premise is given only implicitly. To be considered as enthymematic deductive arguments, inductive reasoning would be considered to always rely on non-expressed premises that make them deductively valid. One such suitable premise could maintain that knowledge gathered from careful, repeated observation of a limited number of cases allows the general acceptance of all similar cases. The inference from premises and the conclusion in inductive enthymematic arguments can follow deductively given this additional premise.

For example, Vorobej (2006, p. 285) discusses the inference from a limited number of yellow duck sightings in a pond to the conclusion that all ducks in the pond must be yellow. When a generic inductive premise is added to this argument, it can indeed be considered to be valid. This does not mean, of course, that the conclusion is true and all ducks in the pond really are yellow. It just makes an argument to that effect valid in a broad sense. Similarly, there are characteristic premises for arguments by analogy (Walton, Reed & Macagno 2008, pp. 55–60) and inferences to the best explanation (Walton, Reed & Macagno 2008, p. 171) that, when added to the other premises of an enthymematic argument, would guarantee the truth of the conclusion in case all premises were true.

This enthymematic treatment of non-deductive arguments can raise concern. It could seem like this broader understanding gives validity away as a *cheap* property (Feldman 2014, pp. 163–165) of arguments. Assume just a few additional premises even to the most outlandish of arguments, and validity ye shall receive. But this would be an exaggeration and a caricature. If one were to include implausible premises in order to achieve this kind of validity, it would easily show in the standard form of the reconstructed arguments. And the argumentation schemes available for non-deductive reasoning provide strong criteria for an argument to be recognised as an instance of a scheme.

A second way to make plausible the assumption that arguments in the models are valid is to allow *defaults* as rules of inferences (Reiter 1980; Straßer & Antonelli 2019, §3.3). While the enthymematic approach adds implicit premises, the default logic approach enlarges the body of valid inference rules beyond those that are valid in the strict sense. For example, inductive or abductive rules of inferences that can not ensure validity in the strict sense could provide validity in the default sense. Since rules of inferences are not modelled in dialectical structure models – every argument is assumed to be valid – this approach provides an alternative assumption to make plausible the validity of a wider range of arguments in these models.

The benefit of these interpretations is that dialectical structure models are plausible representations of a wide range of argumentation in humans, such as in politics or science (where strict deductive reasoning is scarce). In my mind, it is a plausible treatment of validity independent of agent-based modelling – one that has been maintained by serious logicians – but if it is deemed too demanding it could be seen as an idealising assumption that dialectical structure models make. Any framework for agent-based modelling will make idealising assumptions, and these assumptions on validity could be considered as one such idealising assumption.

2.2.2 Debates, dialectical structures, and argument maps

The theory of dialectical structures represents debates as concatenations of arguments. Since arguments are understood in terms of implication relations, this means that the theory studies debates in terms of concatenations of implication relations. This technical understanding of a debate does not need to entail antagonistic controversy, or even multiple participants. An agent that reasons on its own and collects several arguments can be said to have a debate in this technical sense.

Figure 2.3 shows an example for a concatenation of two arguments. In the figure, the debate is represented by a propositional formula, and so it is convenient to call this the *propositional representation* of a debate.

$$\underbrace{((p_1 \wedge \neg p_2) \Rightarrow p_3)}_{\text{Argument 1}} \quad \wedge \quad \underbrace{((p_3 \wedge p_4 \wedge p_5) \Rightarrow p_6)}_{\text{Argument 2}}$$

Figure 2.3: Example for the propositional representation of a debate with two arguments

The arguments in this example are not just concatenated, they are also related to each other. The theory of dialectical structures defines two relations that can hold between a pair of arguments. These relations are known as defeat and support (Betz 2009, pp. 288–289). An argument a_1 defeats another a_2 just in case the conclusion of a_1 is equivalent to the negation of one of the premises in a_2. An argument a_1 supports another a_2 just in case the conclusion of a_1 is equivalent to one of its premises. These two relations do not exhaust any possible way in which two arguments could be related to each other, but several other plausible candidates can be reduced to these two relations (Betz 2009, fn. 10). Betz (2009) originally used the term "attack" to denote the defeat relation, but I prefer the term "defeat" in this thesis to avoid confusion with the argumentation strategy that is also called "attack" (to be discussed in Section 2.4.1.2).

The arguments that meet the defeat and support relations can be written as sets of argument pairs. In Figure 2.3, the set of argument pairs that meet the support relation would be {(Argument 1, Argument 2)}, and the defeat relation would be represented by the empty set. Together with the set of arguments, these sets of defeat and support relations form a tuple, and this tuple is called a *dialectical structure*.

Definition 1 (Dialectical structure, Betz 2009, p. 288). Let T be a set of arguments, A the set of pairs from these arguments for which the defeat relation holds, and U the pairs for which the support relation holds. Then $\tau := \langle T, A, U \rangle$ is called a *dialectical structure*.

Let the representation $\langle T, A, U \rangle$ be known as the *tuple representation* of a debate. Since the dialectical relations that are defined in the theory of dialectical structures can be readily obtained from the arguments themselves, the

tuple representation of a dialectical structure can be easily constructed from its propositional representation.

The tuple representation of a debate indicates how every dialectical structure instantiates an argumentation framework, or, more precisely, an argumentation framework that is both structured and bipolar. The framework in Figure 2.3 with its two arguments (a_1 and a_2) is described by the argument set $T = \{a_1, a_2\}$. No arguments defeat each other, and so $A = \emptyset$, but a_1 supports a_2 and so the pair (a_1, a_2) would be the only member for U. The resulting tuple representation would be

$$\tau = \langle \{a_1, a_2\}, \{\}, \{(a_1, a_2)\} \rangle.$$

Beside the propositional and the tuple representation, there is also a *graph representation* for dialectical structures. This representation uses a two-relation graph to plot all arguments put forward in the debate as well as the defeat and support relations among them. By convention, support relations are represented by solid edges, or $a_1 \longrightarrow a_2$, and defeats are represented by dashed edges, or $a_1 \dashrightarrow a_2$. The graph representation can also be called an *argument map*, and I will use this term as a synonym for a dialectical structure in this thesis. The graph representation for Figure 2.3 is plotted in Figure 2.4.

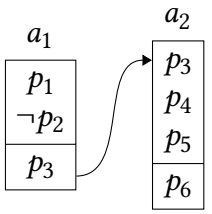


Figure 2.4: Graph representation of Figure 2.3. Arguments are represented as nodes with edges between them expressing either a defeat or support relation. Individual arguments form a two-part node, where the upper part collects the premises and the lower part the conclusion.

These three different representations of dialectical structures serve different purposes. The graph representation is most useful for human readers. The tuple representation is useful to see that dialectical structures instantiate argumentation frameworks. The propositional representation is useful in two ways. First, it illustrates how dialectical structures are actually implemented in the Python implementation for agent-based models developed together with this thesis (see Appendix B). Second, the propositional representation is useful to understand the satisfiability constraint that governs the composition of dialectical structures.

In logic, a formula is said to be "satisfiable" if its variables can be assigned to truth values so that the whole formula evaluates to what is known as the designated truth value. In this thesis, I will assume that the formulas are interpreted in standard propositional logic, which means that the designated truth value is

simply the value True. The satisfiability constraint then demands that a dialectical structure must have a propositional representation that is satisfiable, that is, there must be at least one assignment of truth values to the premises and conclusions such that the concatenation of all arguments evaluates to True. Consider the minimal example with just one argument,

$$(p_0 \wedge \neg p_1) \Rightarrow p_2,$$

which is satisfiable since there is at least one assignment to truth values under which the formula evaluates to True. In fact, there is not just this one but seven such assignments. Here is one such assignment:

$$\{p_0 \rightarrow \text{True}, p_1 \rightarrow \text{False}, p_2 \rightarrow \text{True}\}$$

Any assignment of variables that makes the implication relation expressed by a single argument True attests the validity of the argument. This is because any assignment of such variables will support the claim that, if the premises were true, so must be the conclusion. The satisfiability constraint can thus also be expressed in terms of validity: a dialectical structure can only contain such arguments so that it is possible to simultaneously attest the validity of all arguments.

The processes for the composition of argument maps that I discuss in this thesis all ensure this constraint. But not any hypothetical combination of arguments is an admissible dialectical structure. This is because there are combinations of arguments for which the validity of arguments can not be simultaneously maintained. Consider Figure 2.5. In this example, there is no possible assignment of values True and False so that all individual implication relations simultaneously evaluate to True – although the removal of any of the arguments would make such an assignment possible.

$$(p_1 \Rightarrow p_2) \wedge (p_2 \Rightarrow \neg p_1) \wedge$$
$$(\neg p_1 \Rightarrow p_2) \wedge (p_2 \Rightarrow p_1)$$

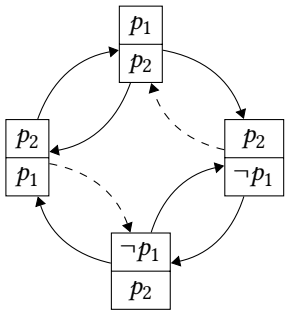


Figure 2.5: A hypothetical circular dialectical structure in propositional form (left) and graph structure (right). This structure can not be realised in an agent-based model because it is impossible to jointly attest validity for all involved arguments in standard propositional logic.

The satisfiability constraint is not the only constraint that governs the composition of argument maps. First, regarding the micro-structure of arguments, arguments can only use propositions as premises and conclusions that are currently available in the sentence pool. To prohibit trivial redundancies, neither

a proposition nor its negation must be the conclusion of an argument if it is already chosen as a premise. And some types of dialectical structure models make further demands of the introduced arguments. For example, the type 1a model (introduced in Section 2.4.1) has a constraint of premise set uniqueness: when a set of premises has been used for one argument, it can not be used for another argument.

Until now I have considered only minimal examples for dialectical structures as well as some constraints for argument map composition that can be explained along the micro-structure of arguments. With this understanding of the micro-structure of arguments, we can proceed to the macro-structure of dialectical structures, and more realistic examples for argument maps that are used in agent-based debate models discussed in this thesis. One example is shown in Figure 2.6.

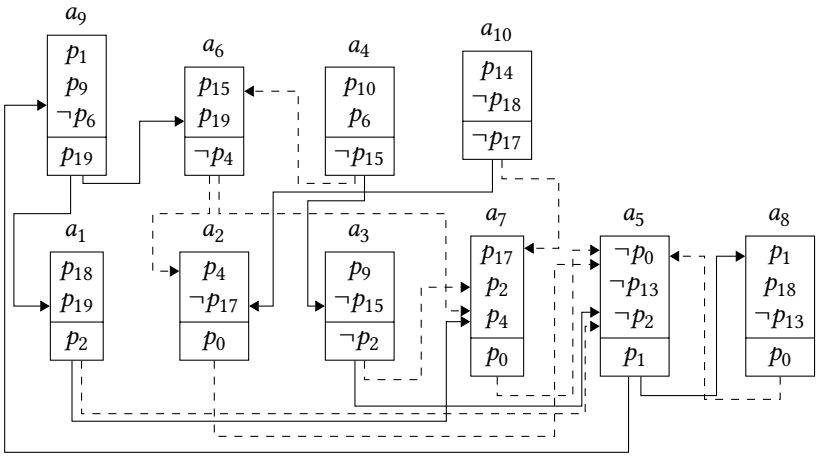


Figure 2.6: Illustration of a dialectical structure with ten arguments

The macro-structure of this argument map might appear unassuming but it is actually quite specific. For one, note that each argument in the graph stands in at least one support or defeat relation to another argument. Second, note that the argument map is hierarchical: there is a group of arguments, displayed on the lower level of Figure 2.6, that lead to conclusions p_0, p_1, p_2. Call these propositions *key sentences* of the debate. There is a second group of arguments, displayed in the upper level of the map, that lead to conclusions which are all premises in arguments on the lower tier. The composition of this map follows an algorithm that ensures this tree-like growth (Betz, Chekan & Mchedlidze 2021, §3).

Dialectical structures do not necessarily have the abundance of relations or the hierarchical order of tree-like argument maps. By contrast, Figure 2.7 shows an argument map that has the same number of arguments as the previous one, but with entirely random relations.

This concludes my introductory review of argument micro-structure and debate macro-structure. In the next subsection, I review the beliefs that agents adopt in light of arguments and dialectical structures.

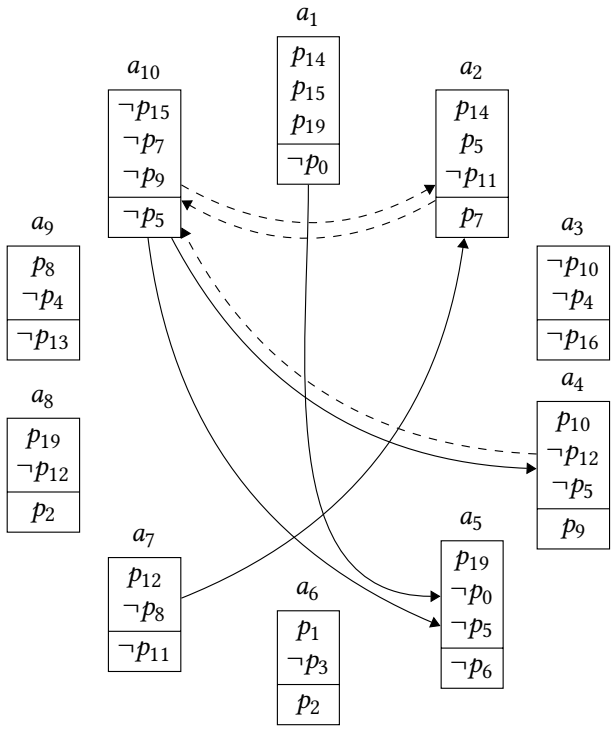


Figure 2.7: A dialectical structure with 10 arguments of random order and with scarce relations. As before, supports between arguments are visualised with solid edges, and defeats through dashed edges.

2.2.3 Agents and their beliefs

For most dialectical structures there are multiple truth value assignments that assert the validity of all presented arguments, and the simulation procedures ensure that, despite changes to the debate, at least one such assignment exists. Agents in dialectical structure models must hold beliefs that reflect such a validity-respecting assignment. These are fundamental assumptions of all dialectical structure models: agents are only confronted with arguments that can be hold valid simultaneously, debates can only advance through the introduction of arguments that maintain this satisfiability, but agents also must hold beliefs that accept the validity of all presented arguments. Should a new argument reveal a validity violation in the beliefs of any agent, that agent must update to beliefs that respect the validity.

The theory of dialectical structures represents agents' beliefs through mappings. These mappings, from the sentence pool to truth values, are known as *positions* (Betz 2013, pp. 34–36). Throughout this thesis, I will use the terms *position* and *belief system* interchangeably. All agents are part of the same debate forum, and they have access to the same set of atomic propositional variables that are being discussed. Agents can assign discrete truth values True and False to the sentences under discussion. In the type 1b model variant (Section 2.4.1), they can also assign None. These three choices represent acceptance, rejection and suspension, respectively. A modified model could also study continuous

truth values from the interval $[0, 1]$, but I did not pursue this approach for this thesis. The definition for a position is given in Definition 2, and Figure 2.8 gives an example.

Definition 2 (Position, Betz 2013, pp. 34–36). Let S be the sentence pool of a debate τ and V the set of possible truth values. A position P is a mapping $S \to V$. A position P is *complete* if and only if it assigns a truth value for every element in S, or, in other words, if $\text{domain}(P) = S$. A position is called *partial* otherwise.

$$
\begin{aligned}
p_1 &\to \text{True} \\
p_2 &\to \text{False} \\
p_3 &\to \text{None} \\
&\vdots \\
p_n &\to \ldots
\end{aligned}
$$

Figure 2.8: Abstract example for a belief system

Epistemologists like Friedman (2013) argue that judgement suspension should be recognised as a third discrete attitude by which epistemic agents interact with propositions. In Figure 2.8 this attitude is represented by the None value, although None is not a truth value in standard propositional logic. In the context of this thesis, I have solved this issue by implementing positions that suspend as partial positions, with the propositions being suspended on not present in the agent's belief system. While this implementation is technically sound, for epistemological purposes it can only be a first approximation to judgement suspension as a state in its own right. Most notably, this implementation can not distinguish between ignorance and suspension. In future work, I aim to provide a more satisfactory implementation of suspension. This work will rely on the ongoing research into the different kinds of suspension (Zinke 2021) and how they are best implemented in agent-based models (Schuster 2022).

When no arguments are present in a debate, there are m^n possible complete positions for a sentence pool of n sentences and m truth values. The experiments reported in this dissertation are performed on pools with a maximum extension of $n = 20$ sentences. This sentence pool extension is determined through computational limitations that might very well be extended in the future. When all sentences are known to them, agents in this thesis can thus adopt one of 2^{20} positions when they can assign either True or False to a sentence, and 3^{20} when they can have partial positions and may also assign None.

The absence of arguments gives agents maximum liberty in selecting among the theoretically possible beliefs. The introduction of arguments constraints this belief choice and drives belief dynamics. For a minimal example, consider how the beliefs $\{p_0 \to \text{True}, p_1 \to \text{True}, p_2 \to \text{False}\}$ becomes unavailable in case the argument $(p_0 \wedge p_1) \Rightarrow p_2$ were to be added to the debate. Any position with these beliefs would not attest the validity of that argument, and

would therefore be inadmissible. I explore the belief dynamics in more detail in Section 2.4.

To verify epistemically rational behaviour in dialectical structure models, agents are not allowed to adopt just any belief system. In Section 1.3, I identified two main epistemic rationality constraints, coherence and responsiveness. Agents with coherent beliefs do not accept flat-out contradictions: they accept exactly one of $p \rightarrow$ True, $p \rightarrow$ False, or $p \rightarrow$ None. Coherence also implies cohering with the arguments known to the agent, and so agents' beliefs must attest to the validity of all presented arguments. I interpret responsiveness to mean that agents must respond to any argument introduction and must consider every available argument in their search for admissible beliefs. They are also responsive by following the arguments where they lead them. If an agent accepts all premises of an argument, it also accepts its conclusion. Call this the "closedness" of the agents' beliefs. Within these constraints, agents are free to select beliefs. This includes the possibility that multiple agents share the exact same beliefs, and it is also possible that a belief system is never maintained by an agent in the population even though it would be perfectly admissible.

The set of all complete belief systems that agents could adopt in light of a dialectical structure, while assigning either True or False to each proposition, is called the *space of coherent and complete positions*, or SCCP (Betz 2013, pp. 39–41). Figure 2.9 visualises the SCCP for a debate with five elements in its sentence pool, $p_1, ..., p_5$, and two arguments. Each admissible complete belief system is plotted in a separate node in the right panel. Two positions are connected through an edge just in case they differ in exactly one truth value assignment. The nodes are labelled by the bit-array of the respective beliefs. The first digit indicates belief in the first proposition (p_1), the second digit in the second proposition (p_2), etc. For example, the node 00100 denotes acceptance of p_3 and rejection of all other propositions. Note that the arguments in the left panel (A) have reduced the theoretically available $2^5 = 32$ belief systems to 26 that remain acceptable, all of which are plotted in the right panel (B).

The SCCP visualises the belief systems that are available to the agents, but it's not a visualisation of which beliefs the agents actually choose. Multiple agents can select the same beliefs available in the SCCP, and often times many possible beliefs are not upheld by any of the agents. The SCCP is a graph, though a quite distinct one from the graph representation of an argument map. Argument maps plot relations between arguments, while the SCCP is best interpreted as plotting a "belief space" that agents traverse as they seek to form validity-respecting beliefs about the sentences under discussion.

From a technical point of view, dialectical structures pose Boolean satisfiability problems, or SAT. As I said in Section 2.2.2, every dialectical structure can be represented by a Boolean formula. The belief systems that agents hold are interpretations under which this Boolean formula evaluates to the designated truth value, which is equivalent to holding beliefs that respect the validity of all arguments. When agents need to change their beliefs, they seek to keep as many previous beliefs as possible. This optimisation task is known as the

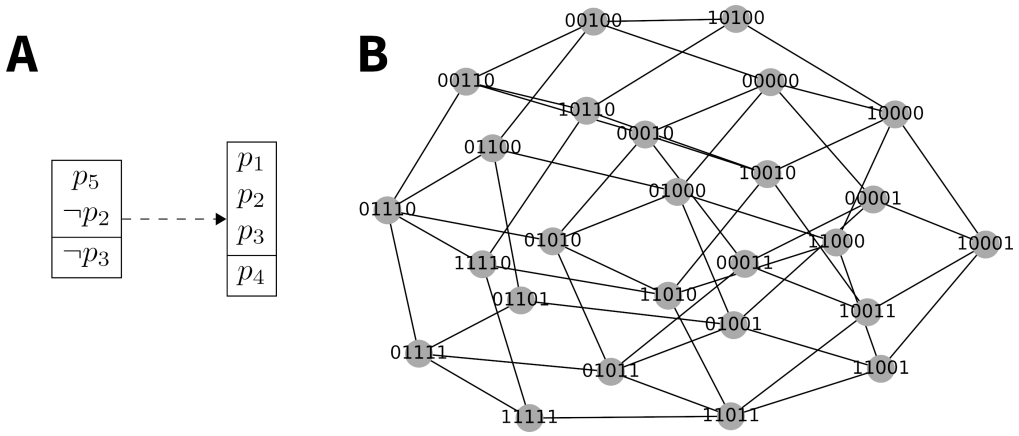


Figure 2.9: All complete, validity-respecting beliefs (B) for an argument map with two arguments (A). Figure from Kopecky (2022). There is a code recipe for the generation of this Figure in Appendix B, page 164.

MaxSAT problem. And finally, the number of interpretations that make the Boolean formula true, which is equal to the number of nodes in the SCCP, is known as the #SAT problem. There exist efficient algorithms for all of these three problems, and these algorithms are used in the implementation of dialectical structure models (see Appendix B).

I have now described the three basic objects of agent-based dialectical structure models, arguments, argument maps, and belief systems. The next section reviews some established measures that can be applied to these objects, and it also introduces some new ones (Section 2.3). Section 2.4 then discusses the dynamics of arguments and beliefs in three different types of dialectical structure models. I conclude this chapter with a brief comparison to other agent-based modelling approaches in social epistemology (Section 2.5).

2.3 Measure what can be measured: Agreement, belief polarisation, opinion diversity, and information entropy

In the preceding section I reviewed the key objects that the theory of dialectical structures describes: arguments and debates, positions and the "belief space" that agents traverse in search of admissible belief systems. Before looking into the dynamics of these objects, in this section I briefly review some established measures for these objects and then introduce some new ones.

2.3.1 Agreement and disagreement

Agents sometimes hold beliefs that are far removed from the beliefs of others, but other times their beliefs make them close neighbours. A distance measure

is a way to quantify this relation. The two distance measures used in this dissertation are based on the sentence-wise comparison of truth value assignments in the positions of two agents. These distance measures are essential to the measurement of (dis-)agreement, opinion diversity and belief polarisation and they are also a fundamental ingredient in the clustering of belief systems.

For agents that assign truth values to the same set of propositions and who either accept or reject on all of them, but do not suspend, the simplest distance measure is the Hamming distance. It simply equals the number of sentences that the agents evaluate differently.

Definition 3 (Hamming distance). Let p and q be two complete positions on a dialectical structure with sentence pool S. For any sentence $s \in S$, $p(s)$ yields the truth value that p assigns to s. Then the Hamming distance is defined as:

$$\mathrm{HD}(p, q) := |\{p(s) \neq q(s) | s \in S\}|$$

The Hamming distance is not robustly meaningful in absolute terms. It makes a lot of difference that $\mathrm{HD}(p, q) = 5$ whether the two positions p and q debate 10 or 100 propositions. Normalising the distance gives it a more universal meaning. Let the size of the sentence pool be given as n. The normalised Hamming distance is then given as $\mathrm{HD}_n(p, q) = \mathrm{HD}(p, q)/n$. Note that n, the size of the sentence pool, is also the maximum Hamming distance between two positions, and so this normalisation is a normalisation onto the range $[0, 1]$.

The Hamming distance is suitable for positions of equal domain, particularly for complete positions on the same debate. For partial positions, or when there are more than two truth values, a suitable generalisation of the Hamming distance is given by the edit distance.

Definition 4 (Edit distance). Let p be a position with domain S_p and q a position with domain S_q. Let w be a weighting function with

- w(switch) the weight associated with switching a truth value
- w(add) the weight associated with adding a truth value
- w(rem) the weight associated with removing a truth value.

Then the edit distance ED is defined as:

$$\begin{aligned}\mathrm{ED}(p, q) := {} & w(\text{switch})|\{p(s) \neq q(s) | s \in S_p \cap S_q\}| \\ & + w(\text{add})|\{s \notin S_p | s \in S_q\}| \\ & + w(\text{rem})|\{s \in S_p | s \notin S_q\}|\end{aligned}$$

The three operations allowed in ED, 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), are equivalent to Gärdenfors's "kinds

of belief changes" (1992, p. 3). As an example, consider how the edit distance between the following two positions is 3:

Position 1	Position 2
$p_1 \to$ True	$p_1 \to$ True
$p_2 \to$ None	$p_2 \to$ False
$p_3 \to$ True	$p_3 \to$ False
$p_4 \to$ False	$p_4 \to$ False
$p_5 \to$ True	$p_5 \to$ None

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. Whether these operations should be weighted differently is a worthwhile question, but one that lies outside the scope of this thesis.

Just as the Hamming distance, the edit distance is also not robustly meaningful in absolute terms. It makes a lot of difference that $\mathrm{ED}(x, y) = 5$ whether the two positions x and y debate 10 or 100 propositions. Consider that the maximum edit distance between two belief systems equals the number of sentences under discussion $n = |S_p \cup S_q|$ multiplied by the weight of the most expensive operation. The edit distance can be normalised to this number:

$$\mathrm{ED}_n(p, q) = \frac{\mathrm{ED}(p, q)}{n \max(\{w(\text{switch}), w(\text{add}), w(\text{rem})\})}$$

In the special case that $w(\text{switch}) = w(\text{add}) = w(\text{rem}) = 1$, the edit distance can be normalised to the union of the positions' ranges, or $|x \cup y|$:

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

If (but only if) all operations are equally costly, the normalised edit distance is a variant of the Jaccard distance, and reduces to the normalised Hamming distance if the positions are complete.

The Hamming and edit distance measures indicate the disagreement between any pair of agent. These pairwise differences can be aggregated to a population-wide mean level of disagreement and agreement. The most straightforward such aggregation is averaging. The one-complement to this averaged distance is introduced by Betz (2013, p. 39) as the measure of population-wide mean agreement.

Definition 5 (Population-wide mean agreement, from Betz 2013, p. 39). Let δ be a distance measure between two agents, such as the normalised Hamming distance of their belief systems, A the population of agents, and $\binom{A}{2}$ the unique pairs in A. Then the population-wide mean agreement is given as:

$$\text{agreement}(A) := \frac{1}{\left|\binom{A}{2}\right|} \sum_{(x,y)\in\binom{A}{2}} 1 - \delta(x, y)$$

Note that the Hamming and edit distances are not the only distance measures that can be defined on dialectical structures. The measures of agreement, diversity, and polarisation are valid for any measure δ that can yield a normalised indication of the difference in belief for any pair of agents.

Normalised distance measures are extremely versatile base measures. For one, they indicate the level of agreement and disagreement in a population. They can also be used directly for some belief polarisation measures, and they are necessary for opinion clustering, which in turn is the base for all opinion diversity and some further belief polarisation measures. I now turn to these more complex measures.

2.3.2 Belief-based clustering

Clustering is the art and science of finding structure and patterns in situations that appear involved and complex at first. Applied to dialectical structures, clustering helps in grouping agents and their positions together: agents with similar beliefs should be in one cluster, and those with very different beliefs should be in different clusters. This cluster analysis helps in determining the belief polarisation and opinion diversity of a group of agents – because, as it turns out, polarisation and diversity are very much concepts related to groups.

Clustering can be endogenous or exogenous (Bramson et al. 2016, pp. 87–88). An endogenous measure arises from the data itself, such as from the difference measures between the positions of agents. Exogenous measures, by contrast, are obtained from characteristics external to the data at hand. When it comes to differences of opinion, the opinion of agents would be an endogenous data source for clustering. Demographic properties, such as the city of residence or age group of the agent, would be an exogenous source antecedent to the difference of opinion itself. In agent-based modelling, we often have to rely on endogenous clustering as agents are only known through their beliefs. A clustering algorithm delivers the role of endogenous structuring.

Clustering algorithms can be simple and straightforward. Imagine that a dialectical structure would describe a debate involving a proposition p and that our interest would solely rest on p. Then we could cluster the beliefs of agents into three groups, based on whether they accept, reject, or suspend on p. But this technique does not generalise well. For the general case, consider a population of five agents with belief systems formed by responding to a dialectical structure. The pairwise distances of these agents can be collected in a quadratic matrix like in Figure 2.10. The cell in position (i, j) in such a matrix would show the difference between agents i and j. Assuming a symmetric distance, this matrix would be symmetrical.

How should the beliefs on this matrix be clustered? Unlike the simple case in which we care just about p, it is not at all obvious from the data. The good news is that state-of-the-art clustering algorithms can be used for partitioning belief systems on dialectical structures. The bad news is that they do not work straight out of the box.

$$\begin{bmatrix} 0.0 & 0.45 & 0.3 & 0.5 & 0.6 \\ 0.45 & 0.0 & 0.65 & 0.55 & 0.35 \\ 0.3 & 0.65 & 0.0 & 0.6 & 0.6 \\ 0.5 & 0.55 & 0.6 & 0.0 & 0.3 \\ 0.6 & 0.35 & 0.6 & 0.3 & 0.0 \end{bmatrix}$$

Figure 2.10: A matrix of normalised, symmetric differences for five agents. In the matrix, the cell (i, j) shows the normalised difference between agents i and j.

Two clustering algorithms turn out to be particularly useful for dialectical structure models, the Leiden algorithm (Traag, Waltman & van Eck 2019) and affinity propagation (Frey & Dueck 2007). The underlying mechanisms for both algorithms differ: Leiden seeks clusters that have maximally strong ties between members – in other words, it optimises the *modularity* of the population. Affinity propagation seeks a set of belief systems that are particularly useful to characterise the beliefs in the population. Those serve as cluster centres and are known as *exemplars*. Each and every agent from the population is then assigned to the exemplar it has the most affinity to.

The distance matrix from Figure 2.10 is not suitable as input to the Leiden and affinity propagation clustering algorithms. First, the two algorithms expect to be given *adjacency*, or "similarity", instead of distance matrices. Second, the two algorithms turn out to work best for sparse adjacency matrices. This is because the algorithms are designed for social networks, and social networks are often sparse and not complete. Consider, for example, how you are ignorant of most other users on a digital social network.

Distance matrices can be transformed to sparse adjacency matrices by first scaling all elements and then filtering all adjacency values that are below a threshold. There is no natural choice of scalar or filter. I found that using a scalar of either e^{-2x} or e^{-4x} and a threshold of 0.2 to be useful for both algorithms. Figure 2.11 shows how the normalised differences in pairs of agents can be transformed and filtered to adjacency values, depending on different scalars. The x axis shows the distance obtained from a normalised distance measure, such as the distances shown in the matrix in Figure 2.10. The y axis shows the resulting scaled adjacency score that will be used for clustering. Scaled similarities below 0.2 are always filtered and set to 0 in order to obtain a sparse matrix. In the case of e^{-4x}, all agents with a higher normalised distance of 0.4 are not considered to be adjacent, and their adjacency score is set to 0 in the resulting adjacency matrix. For debates involving 20 propositions, this means that any two agents are only considered adjacent when they agree on at least 12 propositions. In the case of e^{-2x}, only agents with a normalised distance greater than 0.8 are not considered adjacent. Sharing this level of agreement does not necessarily imply that these agents belong to the same cluster. Agents in a cluster are adjacent to each other, but adjacent agents are not necessarily in the same cluster.

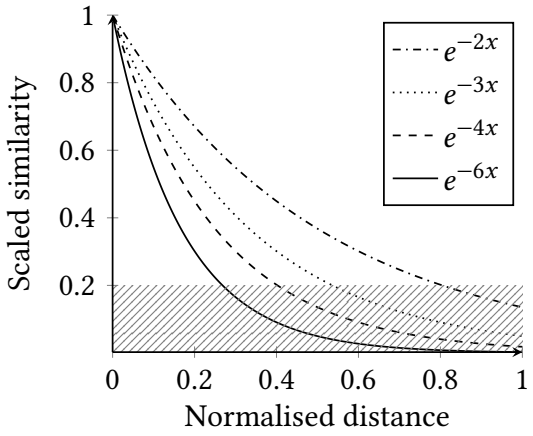


Figure 2.11: Scaling and filtering for normalised distances for the community structuring algorithms. Values in the shaded area are flattened to 0. Note that a normalised distance of 0 is always transformed to a adjacency of 1.

Figure 2.12 shows the scaled adjacency matrix that results from the application of the scalar e^{-4x} and a filter of 0.2. The clustering algorithms can then be applied to this matrix. The clustering obtained from applying the Leiden algorithm to this adjacency matrix is shown in Figure 2.13.

$$\begin{bmatrix} 1.0 & 0.0 & 0.3 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.25 \\ 0.3 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.3 \\ 0.0 & 0.25 & 0.0 & 0.3 & 1.0 \end{bmatrix}$$

Figure 2.12: Adjacency matrix obtained from the distance matrix in Figure 2.10 with a scalar of $e^{-4\delta(i,j)}$ and a filter of 0.2.

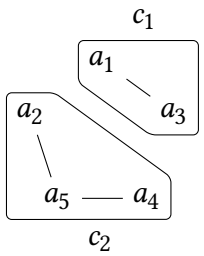


Figure 2.13: Leiden clustering with clusters c_1 and c_2 obtained from the adjacency matrix in Figure 2.12.

The clustering yields two clusters, one containing the two agents a_1 and a_3 and the other one containing the agents a_2, a_4 and a_5. This is reflective of their distances shown in Figure 2.10. It might seem odd though that agents a_2 and a_4 are grouped in one cluster, given that their normalised difference of 0.55 is comparatively high. In particular, both agents have a lower distance to a_1 (0.45 and 0.5, respectively) without sharing a cluster with a_1. But consider how they also have a considerable lower distance to a common neighbour a_5 (0.35 and 0.3, respectively). This agent bridges their difference and makes for a plausible

grouping. The pairs of agents connected through an edge in Figure 2.13 have a disagreement of either 0.3 or 0.35, the lowest two values in the population.

Some polarisation and all diversity measures rely on antecedent clusterings, but they make different usage of clusters. Measures of belief polarisation track the internal agreement in each cluster and the external disagreement to other clusters. Polarisation is an aggregation of differences adjusted to cluster membership. Opinion diversity is not so much about the distances within and between clusters, but more about the shape, size, and number of clusters. When diversity is understood in terms of richness, it gives the number of clusters. Understood in terms of heterogeneity, diversity indicates the likelihood of two randomly drawn agents to belong to different clusters.

At this point you might ask: how reliable are the clusterings we obtain from Leiden and affinity propagation, given that they have not previously been applied to dialectical structures? One way to ensure their reliability is to verify that they do not return erratic measures throughout a simulation run. I have used the adjusted Rand index (Hubert & Arabie 1985) to compare the clusterings of agents in adjacent simulation steps. The assumption is that agents should roughly remain in the same cluster as in the previous simulation step, though some movement should be allowed. This would translate into a high, but not maximal adjusted Rand index. Figure 2.14 shows data that verifies this desired behaviour.

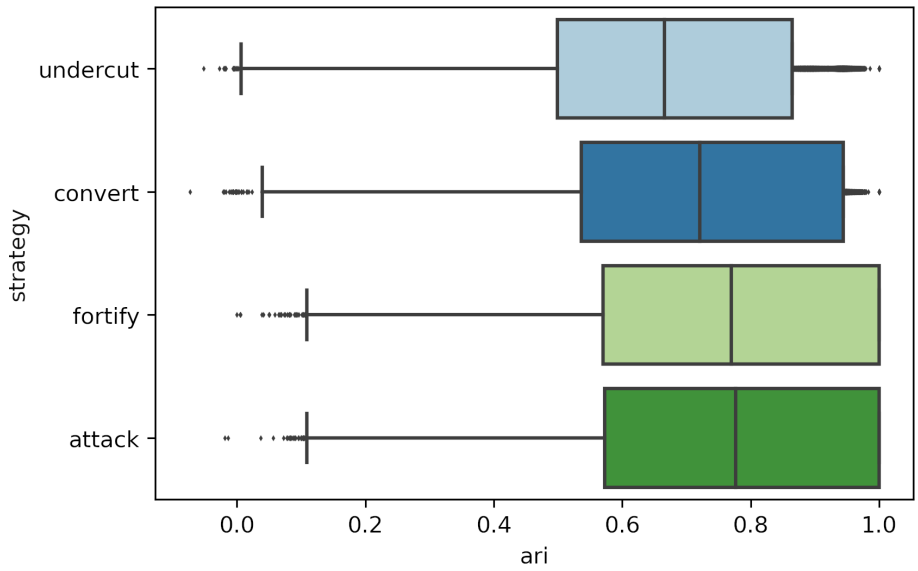


Figure 2.14: Reliability of Leiden clusterings for adjacent simulation steps in the type 1a model from Chapter 3 as estimated by the adjusted Rand index. Figure from Kopecky (2022).

I have also ensured clustering reliability by comparing the measures obtained through cluster-based measures with measures that do not rely on clustering. This is indeed possible for measures of belief polarisation by comparing the group divergence measures with those obtained in statistical dispersion. As

I show in Chapter 3, results obtained from these two measures are very similar indeed. Let us turn to the discussion of these polarisation measures now.

2.3.3 Belief polarisation

In Section 1.4, I said that polarisation is the formation of clusters that converge internally but move away from other, likewise internally convergent, clusters (Esteban & Ray 1994, p. 824). Two types of polarisation in humans are issue polarisation, or the belief-based clustering of opinions, and affective polarisation, or the clustering of groups based on in-group sympathy and out-group antipathy. In this thesis, I study issue polarisation among agents that form beliefs in light of a dialectical structure.

I measure polarisation among agents in dialectical structure models with three measures from Bramson et al. (2017): group divergence, group consensus, and dispersion. The first two measures assume a partition of beliefs into clusters. 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 the intuitive understanding of polarisation that I have mentioned before: when this happens, groups become both more internally alike and externally alien.

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 6) and group consensus (Definition 7) lie in this interval as well.

Definition 6 (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 7 (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:

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

These two measures rely on an antecedent partitioning into groups, such as those obtained through the two state-of-the-art community structuring algorithms for social networks that I discussed in Section 2.3.2. As discussed there, the clustering algorithms are run on scaled and filtered distance matrices. Using multiple algorithms is one strategy to verify that the obtained clusterings are reliable.

Dispersion is a third polarisation measure from Bramson et al. (2017), but one that does not rely on an antecedent clustering. Dispersion tracks an intuitive idea of how agents and their belief systems can polarise by measuring how agents deviate from a population-wide mean. If these spread out evenly or cluster around one pole, dispersion will be low, but clustering around increasingly distant poles will lead to increased dispersion.

When agents' belief systems are understood in terms of positions towards a debate stage, it is usually impossible to define a population-wide mean. This is because there will often be several, but distant, positions that likewise maximise centrality measures, and graphs on positions often have more than one graph centre.

A way to avoid this is to replace the mean position with the mean distance between all pairs of positions, and then inspect the dispersion of distances around that mean. But one must be careful here to select a polarisation measure that is an *aggregation* of distances – merely interpreting the average distance between pairs of agents as dispersion would lead to a concept of polarisation that is too close to a concept of agreement. Following Bramson et al. (2016, p. 84), I use the standard deviation of pairwise distances as a measure of dispersion (Definition 8).

Definition 8. Dispersion, understood as the standard deviation on the pairwise distances between agents' belief systems. Let δ be the normalised edit distance and A_τ the set of agents at debate stage τ, represented by their positions. $\binom{A_\tau}{2}$ denotes the pairs of agents in the population. With $N = |A_\tau|$, dispersion is defined as:

$$\text{dispersion}(\tau) := 2\sqrt{\frac{1}{N}\sum_{i\in A_\tau}^{N}\left(\frac{1}{N-1}\sum_{j\neq i}^{N-1}\delta(i,j) - \underbrace{\frac{1}{\left|\binom{A_\tau}{2}\right|}\sum_{(x,y)\in A_\tau}\delta(x,y)}_{\text{mean distance (constant)}}\right)^2}$$

In comparison to other polarisation measures, dispersion does not require a computationally intensive clustering, while still approximating, as discussed in Chapter 3, the values obtained from cluster-based polarisation measures. This is helpful to confirm that both the clustering and the obtained polarisation measures are reliable.

A good question to ask is how measuring polarisation is different from measuring agreement and disagreement. To recall from Definition 5 (page 40), Betz (2013) measures disagreement as the *averaged* normalised distance between pairs of agents' belief systems, and agreement as the inverse of this value. The

polarisation measures discussed in this section aggregate from the same pairwise normalised distances between agents, but they do not use the uniform aggregation of the mean. Dispersion tracks the standard deviation of pairwise distances away from the mean distance in the population. Group divergence and group consensus aggregate the distances depending on which opinion cluster the agents belong to. Measuring polarisation thus goes beyond reporting an absence of agreement in a population and further characterises such disagreement in terms of variation and clustering.

2.3.4 Opinion diversity

In ecology, an ecosystem sample can be measured for diversity by calculating the relative frequencies of all species in the sample. As I said in Section 1.5, this implies an interpretation of diversity in terms of *heterogeneity* rather than *richness*. The collected relative frequencies can indicate how likely 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, p. 856; Page 2011, Chapter 2), and I find an analogous measure to be informative about opinion diversity as well.

The Gini–Simpson index is related to other diversity indices such as the Shannon index (Tuomisto 2010). It is an adequate measure for dialectical structure models because it is refined on smaller populations of less than 100 individuals (Tuomisto 2010, p. 854), a common population size for agent-based models. I adjust the definition from Tuomisto (2010, p. 856) to the purpose of agent-based models:

Definition 9 (Gini–Simpson diversity index). Let A be a population of agents and $T = \{t_1, t_2, ..., t_n\}$ 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:

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

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, I find it plausible to cluster agents into types based on their beliefs. In Chapter 5, I use diversity measures based on the affinity propagation clustering algorithm discussed in Section 2.3.2. The diversity values observed in this thesis thus depend on the reliability of the chosen clustering algorithm.

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

Agreement, diversity, and polarisation are all characterisations of the difference of opinion in a group. And so a natural question to ask is how measures of opinion diversity relate to these other measures. Diversity measures characterise populations in terms of how frequently a type is expressed. Polarisation

measures characterise populations differently, through the distances between individuals. 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, pp. 122–128).

There are interesting implications between agreement, diversity and polarisation. Polarisation is maximal when the 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 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 incurs lowering diversity, or "simplification" (Bednar 2021, pp. 3–4). Only in the special case of a population with exactly two far-removed clusters can polarisation and diversity simultaneously reach their maximum. A two-typed population will have both maximal polarisation and maximal diversity when the two clusters have the same number of individuals in them. In general though, the concepts of issue polarisation and opinion diversity diverge.

How exactly a population should be measured for diversity and polarisation can be a matter of dispute. For example, one of the polarisation measures that Bramson et al. give, "size parity" (2016, p. 93), is not obviously a measure of polarisation. This measure is equivalent to the normalised Shannon diversity index (see, e.g., Page 2011, p. 69). The Shannon index is a measure of group sizes, but not of differences among the groups, and is considered as an index of diversity in the literature (Tuomisto 2010). But if it is considered a diversity index it cannot also be a measure of polarisation. For this reason I consider it best to regard group size parity, i.e. the Shannon index, as a diversity index and not as a measure of polarisation.

2.3.5 Inferential density and information entropy

Measures of agreement, polarisation, and diversity are tools to understand the difference of beliefs that agents hold in light of a dialectical structure. There are also two measures that help us understand a dialectical structure itself. These measures are inferential density and normalised information entropy. They give, first, an indication of debate progress in evolving debates, and second, a measure of how much the beliefs of agents are constrained through the arguments in belief choice.

As argument maps grow through the addition of arguments, inferential density is a normalised measure in $[0, 1]$ that indicates the progress of a debate (Betz 2013, p. 44). The introduction of arguments adds inferential constraints to the beliefs that agents can choose. This process reaches a natural end when only one position remains compatible to the validity of all presented arguments. At this point, it does not make sense to introduce further arguments

as agents can not update their position any further. Any other position they would move to would violate the validity of at least one argument. But not all argument introductions contribute equally to reaching this natural end. In fact, both the number of premises and the choice of argumentation strategy influences the number of arguments necessary to obtain a specific point of inferential density (Betz 2013, pp. 47–49). The measure is defined as follows:

Definition 10 (Inferential density, from Betz 2013, p. 44). Let τ be a dialectical structure, n the size of its sentence pool and σ_τ the number of complete and coherent positions on τ. Then the inferential density $D(\tau)$ is defined as

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

Note that 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 sheer number of agents or arguments.

Two points, $D(\tau) = 0$ and $D(\tau) = 1$, are of particular interest. In case no argument has been introduced, there are no constraints on belief choice. Any of 2^n logically possible combinations of beliefs are acceptable, and so, since $\log_2 2^n = n$, it holds that $D(\tau) = (n - \log_2(2^n))/n = 0$. At the other extreme, at $D = 1$, the inferential obligations from the arguments are so tight that only a single validity-respecting belief system remains, and since $\log_2(1) = 0$, then $D(\tau) = (n - 0)/n = 1$. The argument map then predetermines all beliefs. Inferential density thus gives a measure that is minimal when there are no inferential constraints at all but that is maximal when the constraints are maximal.

To summarise, density is not only an indicator of debate progress, but also a measure of the degree of freedom that agents have in selecting their beliefs in light of an argument map. Agents in dialectical structure models only hold beliefs that are compatible to the validity of all presented arguments, and newly introduced arguments always have the potential of making a previously held position inadmissible. Inferential relations can pre-determine the beliefs that agents hold, and inferential density quantifies this influence.

Density is the one-complement of the argument map's *normalised information entropy*, or $H_N(\tau)$, and it holds that $D(\tau) = 1 - H_N(\tau)$. We can thus understand inferential density as determining the amount of inferential information encoded in an argument map. This is a very useful relation, because this further clarifies the somewhat loose sense of an agent's "degree of freedom" in selecting its beliefs. Using information entropy, it is possible to quantify how many choices can agents make freely, on average, before the arguments determine their other 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.

To derive the formal relation between density and 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.

Some, but not all, of the 2^n belief systems will respect the validity of all presented arguments 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 dialectical structure models will only accept validity-respecting beliefs, $p(i)$ will take one of two values:

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

I now use this property of $p(i)$ to transform the normalised entropy $H_N(p)$. Observe that, with $N = 2^n$:

$$H_N(p) = -\sum_i \frac{p(i)\log_b p(i)}{\log_b N} = -\underbrace{\left(\frac{\frac{1}{\sigma_\tau}\log_2\left(\frac{1}{\sigma_\tau}\right)}{\log_2 2^n} + \ldots + \frac{\frac{1}{\sigma_\tau}\log_2\left(\frac{1}{\sigma_\tau}\right)}{\log_2 2^n}\right)}_{\text{repeated } \sigma_\tau \text{ times}}$$

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:

$$H_N(p) = -\frac{\sigma_\tau \frac{1}{\sigma_\tau}\log_2\left(\frac{1}{\sigma_\tau}\right)}{\log_2 2^n}$$

Here it is useful to recall some general properties of the logarithm. First, for all positive real numbers b except $b = 1$, $\log_b b^n = n$. Second, the quotient $\log_b(x/y)$ can also be written as the sum $\log_b x - \log_b y$. And lastly, $\log_b 1 = 0$ for all bases b. These properties of the binary logarithm allow this transformation:

$$H_N(p) = -\frac{\sigma_\tau \frac{1}{\sigma_\tau}\log_2\left(\frac{1}{\sigma_\tau}\right)}{\log_2 2^n} = -\frac{\log_2\left(\frac{1}{\sigma_\tau}\right)}{n} = -\frac{-\log_2 \sigma_\tau}{n} = \frac{\log_2 \sigma_\tau}{n} = 1 - D(\tau)$$

Now we can say that the inferential density of an argument map τ is the one-complement to its normalised entropy, or $D(\tau) = 1 - H_N(\tau)$.

I have now reviewed the fundamental objects in dialectical structure models and how we can measure agreement, opinion diversity and belief polarisation for agents that traverse dialectical structures. We have also learned how argument maps and belief systems are related through inferential density and information entropy. The only piece that is missing now is a description of the changes to arguments and beliefs in dialectical structure models. I now turn to these dynamic aspects of simulation procedure.

2.4 Dynamics and simulation procedure in evolving and synthesised dialectical structures

The theory of dialectical structures offers a rich framework for the design and study of agent-based models. Agent-based dialectical structure models build epistemic environments out of arguments and debates and equip agents with belief systems to interact with this world. This section is about the dynamics of arguments, argument maps, and belief systems in these models. It is about the factors and processes that change the objects described so far in this chapter.

This thesis studies two types of dialectical structure models. The first type, with two subtypes, is closely related to Betz's (2013) earlier study of consensus and veracity evolution. In the base type 1a model, argument maps grow through agents' introduction of arguments, and agents update their beliefs in response to newly introduced arguments. The extended type 1b considers additional factors, such as tree-like rather than random debate growth, ternary positions with the ability to suspend, and an extending instead of a fixed sentence pool. Results obtained from a type 1b model support my case for epistemically rational belief polarisation. In the type 2 dialectical structure model, argument maps do not evolve through introduction of arguments, but are synthesised without the agents being involved. Instead, an algorithm generates tree-like argument maps with a desired number of key propositions and a target inferential density. Populations of agents with admissible positions are then sampled to obtain a desired opinion diversity or belief polarisation among them. Simulations on type 2 models reveal that samples with high opinion diversity have a much higher risk of inconsistent majoritarian judgement aggregation. Table 2.1 summarises the differences among these models.

Table 2.1: Summary of dialectical structure model types.

	Type 1a	Type 1b	Type 2
Debate composition	Evolving	Evolving	Synthesised
Argument relations	Random	Hierarchical	Hierarchical
Agents' stances	Binary	Ternary	Binary
Sentence pool	Constant	Expandable	Constant

2.4.1 Model type 1: Belief dynamics in debates that evolve through incremental additions of arguments

2.4.1.1 Set-up and initialisation

Type 1 dialectical structure models are variations of the original computational model (Betz 2013). While type 1a is quite faithful to the original, type 1b introduces new features: the possibility for agents to withhold judgement, extension of the sentence pool, and tree-like debate growth.

Type 1 models have a public debate forum. Agents are randomly selected to add an argument to this forum. All agents are aware of all introduced arguments and the belief systems of all other agents. Debates start with an empty forum and a pool of n atomic propositions. This sentence pool remains fixed in type 1a models. In type 1b, it 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. In type 1b, a subset of atomic propositions is deemed to contain *key statements* of a debate. This is because type 1b features tree-like, or hierarchical, growth of the argument map (see Section 2.4.1.2). The argument maps in type 1a grow randomly and propositions are not distinguished into key and auxiliary items.

Changes to the debate forum over time are tracked in terms of *debate stages*, referred to by 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 towards this formula. Debates in type 1 models are initialised without any arguments, and so the first debate stage is an empty formula. The general layout of such a formula is given in the example below:

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

Agents are fully aware of all available propositions, but they are not aware of propositions that are not yet introduced into the forum. In type 1a models, agents can either accept a proposition or reject it. In type 1b models, they can also suspend judgement by not assigning a truth value. At model initialisation, agents randomly assign one of the available truth values to each proposition in the current sentence pool, though robustness analyses for both model types also show the effects of different initialisations (see Chapters 3 and 4).

A computer simulation progresses by scheduled events and there are three such events in type 1 dialectical structure models. These are argument introduction, position updating, and sentence pool extension. The last one, the addition of a new sentence, is only ever called in the type 1b model and never in type 1a. When both argument introduction and proposition pool expansion are allowed, they occur with a different chance (9:1 in the type 1b models in this thesis). Argument introduction and proposition pool expansion are always followed by a position updating event. Figure 2.15 gives an overview of a simulation run, the elements of which are explained in the following subsections.

There are dynamic aspects to the rationality criteria that the agents must meet in their beliefs. Coherence, closedness and responsiveness depend not only on agents' beliefs but also on the current debate stage. As the agents have to respond to newly presented arguments, beliefs that have been coherent and/or closed at stage τ_i, can become incoherent and/or not closed at τ_{i+1}. Section 2.4.1.4 describes how agents respond to cases of rationality violations. These criteria describe agents' behaviour towards the forum – but there are also constraints on how the agents shape this forum in argument introductions. These are described next in Section 2.4.1.2.

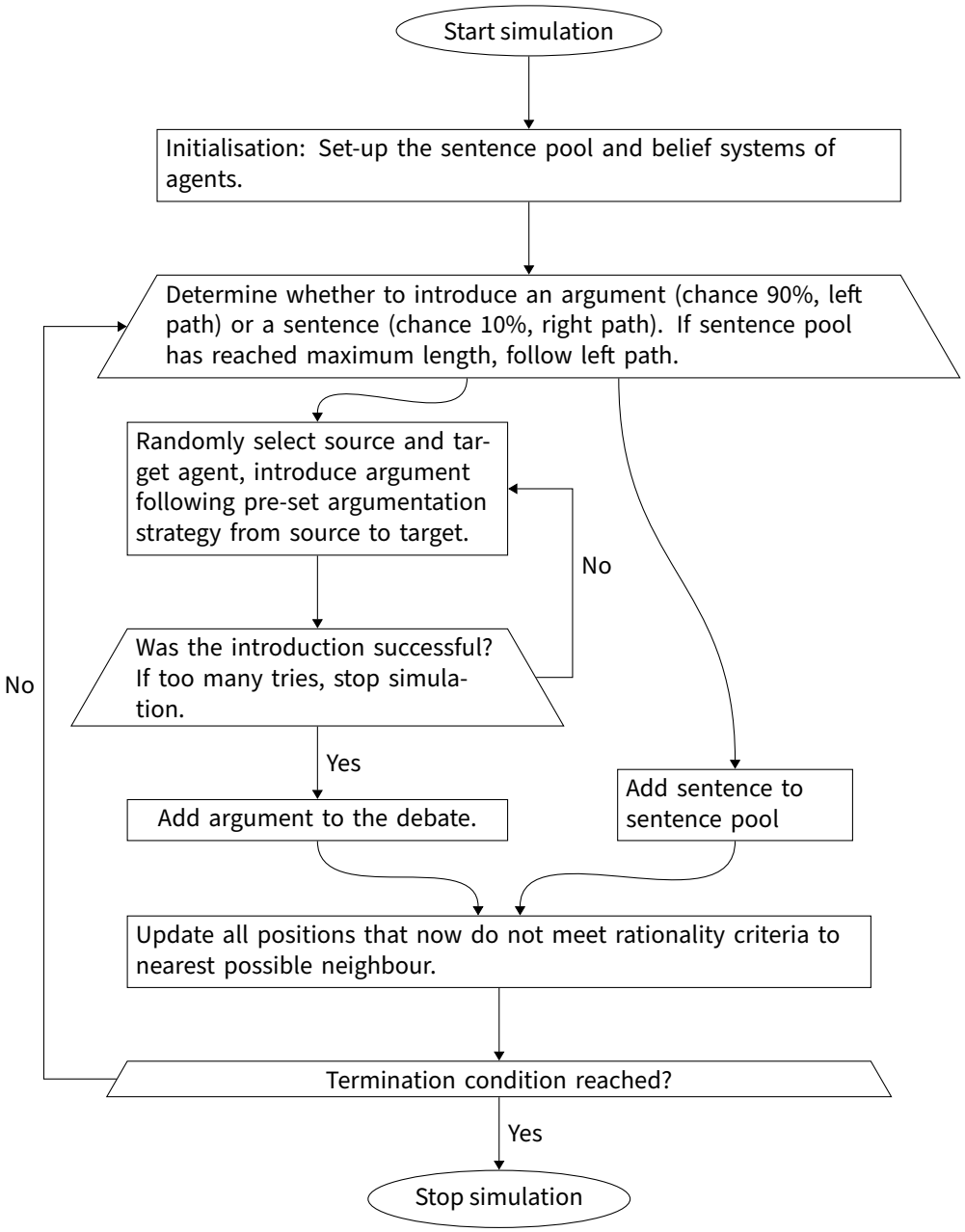


Figure 2.15: Overview of a simulation run for the type 1b model (from Kopecky 2024). Rectangles contain events and decisions are marked by trapeziums. The type 1a model is similar except it does not have the right path (sentence pool extension).

2.4.1.2 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 source and target.

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. The first four are taken from Betz (2013, pp. 93–94), and the last one was added for the purpose of this thesis:

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 we could think about finding evidential support for our beliefs: the agent selects one or more premises about evidence, possibly together with auxiliary premises on general procedures or principles, to support a conclusion it accepts. Change the conclusion to the negation of a belief 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 dialectical structure models they must allow agents to revise their beliefs in accordance with the rationality criteria described earlier in Section 2.2.3.

The argument introduction procedure is different for type 1a and 1b models. Type 1a models grow random argument maps and type 1b grow them hierarchically. So in type 1a models, the source picks premises that fit the argumentation strategy and the target (if present) first and then chooses a conclusion.

In type 1b models, 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. Through implementation of an algorithm from Betz, Chekan & 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 Figure 2.16). The premises of arguments that lead to key statements are then selected as conclusions on the second tier and the same pattern repeats. This hierarchical ordering produces argument maps with a different shape compared to random graphs that do not have discernible key issues.

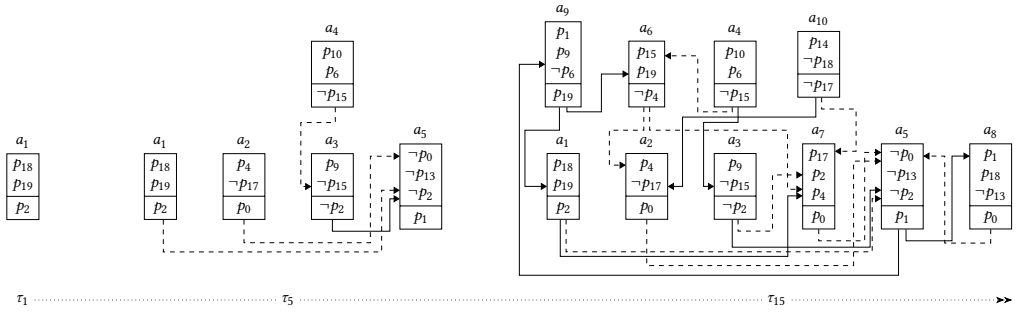


Figure 2.16: Tree-like debate growth with key statements p_0, p_1 and p_2, showing three debate stages τ_1, τ_5 and τ_{15}. (Figure from Kopecky 2024)

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.

2.4.1.3 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. An introduced proposition is available for all subsequent argument introductions.

2.4.1.4 Position updating

Agents revise their beliefs after argument introduction and proposition pool expansion to uphold an epistemically rational outlook on the debate. Belief revision is only triggered in these two ways to isolate the effects of informational influence through argumentation. The model in general would be adjustable to other forms of revision and in particular to agent interaction through social influence.

The updating following the introduction of a new proposition is rather simple: every agent randomly assigns one of True, False, or None (in type 1b) 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 trough 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 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. 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 theoretical works about belief revision. Quine & Ullian (1978, pp. 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, p. 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.

The rationality of closest-coherent updating can also be 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. 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.

The dynamics of type 1 models have a natural end point: at a density of $D = 1$, only a single validity-respecting belief system remains for the agents to choose. If the sentence pool has reached its maximum extension, no further argument can be devised to change the agents' mind. Letting the model run until $D = 1$ can provide some insight indeed (to be discussed later in Section 4.3.7). But following the convention established by Betz (2013, p. 95), most model runs are studied until $D = 0.8$.

2.4.2 Model type 2: Epistemic group decision problems posed by synthetically generated argument maps

Type 1 dialectical structure models study the evolution of debates and how the agents that participate in them change their minds, form consensus, or split into polarised groups. Type 2 models do not look at consecutive debate stages or the dynamics of belief, but instead synthesise a single argument map and a population of agents with static beliefs in response to this argument map. Later, in Chapter 5, I discuss experiments that use this model variant to study epistemic group decision problems. In these problems, argument maps characterise the epistemic circumstances that the groups experience, and the group of agents with different views on these arguments is tasked with finding a shared belief system in response to this argument map. I will ask how successful groups of different opinion diversity and belief polarisation are to form consistent group opinions through sentence-wise majority voting.

Epistemic group decision problems are modelled with 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 population with arbitrary degrees of opinion diversity and belief polarisation, as understood by the measures from Section 2.3. Both sub-processes fundamentally rely on the theory of dialectical structures to describe arguments, belief systems, and what it means that agents accept the validity of all presented arguments.

2.4.2.1 Argument map synthesis

The first sub-process synthesises a collection of arguments that constitutes the basis of the agents' decision problem. The argument collections are synthesised in such a way that there are quite a few beliefs that respect the validity of all presented arguments, resulting in decision problems under epistemic permissiveness. The example I used in the introduction to illustrate independence phenomena (Table 1.1 on page 4) can be seen as a minimal example of this process. Recall that it 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 a substantial amount of arguments. This is why type 2 models generate complex argument maps as opposed to this minimal but illustrative example.

An example of the actual argument synthesis process is pictured in Figure 2.17. 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, a subset of propositions is designated as the key propositions of the debate. These propositions can be imagined to be most central to the decision problem. Arguments 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. With the algorithm from Betz, Chekan & Mchedlidze (2021, §3), arguments are added iteratively to the map until a specified value of inferential density is reached.

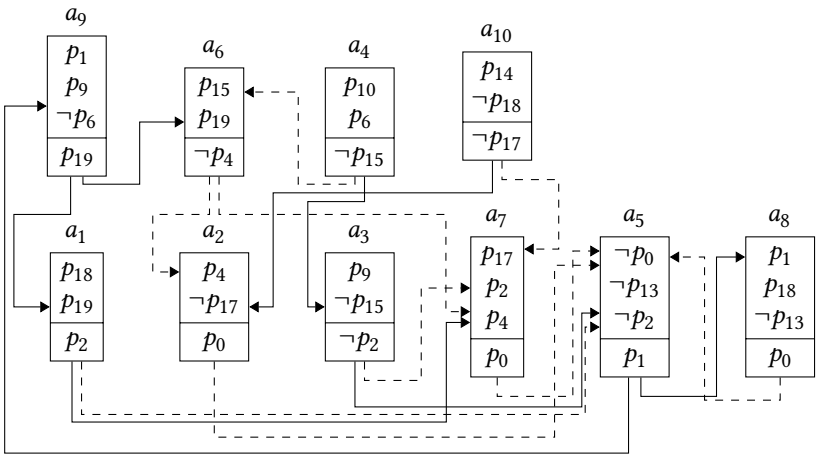


Figure 2.17: Illustration of a synthetically generated tree-like argument map with key statements p_0, p_1, and p_2. As before, support relations are expressed by solid arrows, defeats by dashed ones.

2.4.2.2 Agent population sampling

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

Agents' beliefs are expressed by a mapping from the propositions in all arguments to binary truth values True and False. Every agent assigns a truth value to every proposition from the argument map. As in other dialectical structure models, all agents hold beliefs that respect the validity of all presented arguments. This simplification is necessary to allow for voting without abstention, and it implies that I am modelling quite specific scenarios. As all agents are competent to judge all involved propositions, the model is best interpreted as tracking the decision procedure in agents with relatively large overlap in expertise. Extensions of this model type could track decision procedures in multi-disciplinary groups by allowing suspended judgement in the voting procedure.

From agents with beliefs that are characterised in this way, the model samples groups with a specified degree of diversity or polarisation. Since I 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 from Chapter 5, 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. My search for groups with specific expressions of diversity and polarisation thus has to be strategic. I describe the group sampling algorithms in more detail in the supplementary materials for Chapter 5, but I include a brief summary here. For the diversity variant of the model, I first apply the affinity propagation clustering algorithm (Frey & Dueck 2007) to the collection of all beliefs that respect the validity of the antecedently synthesised map. As I regard membership in these clusters as type expression, I 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, I sample agents following a pyramid scheme of sorts: for a given distance δ, I initially draw a pair of agents with mutual distance δ in their beliefs. I 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 Chapter 5, I present the results from such a simulation experiment.

2.5 The approach in perspective

Computational approaches to belief dynamics in social settings now offer a rich field with diverse approaches from different disciplines and technical frameworks. Models built on the theory of dialectical structures are similar to some of them and distinct from others. My purpose in this section is not to offer a full review of the complete field – there are others who have done so much better. Grim & Singer (2024) and Šešelja (2023) offer overviews of this field, and Šešelja (2022) offers a systematic and in-depth comparison of network epistemology and abstract argumentation models. My review in this section can not be as encompassing or in-depth as these reviews, but I will

devote particular attention to two types of models. The first are those that model belief dynamics through argumentation, similar to dialectical structure models. Models in the second group yield results that are comparable to the results obtained from dialectical structure models in this thesis. However, I will not consider agent-based models in evolutionary game theory (e.g., Bruner 2015) and some other approaches.

In discussing these models, I am pursuing four questions:

1. How do they implement the epistemic environment that agents traverse? What data is available to agents?
2. How are agents' belief systems implemented?
3. How do agents update their beliefs? Do they base their updates on the beliefs of others (social influence) or do they (also) take information from evidential sources (informational influence)?
4. Which phenomena can be studied through the model?

Below, the first two models are interesting because they use models of arguments and argumentation. The other models are interesting because they study similar phenomena that are investigated in this thesis.

It should be noted that the modelling frameworks and techniques discussed in this section are not necessarily mutually exclusive. For example, agents could have beliefs that meet the requirements of Bayesian epistemology but traverse an epistemic environment that is described by an abstract argumentation framework. Or, alternatively, an epistemic landscape. Because these models have different strengths and limitations, for the computational philosopher it seems most recommendable to understand as many aspects of all modelling approaches as possible. In this way we can make informed decisions for which simulation experiments a framework can be used, and for which phenomena a different approach would be better suited.

2.5.1 Abstract argumentation frameworks

Dialectical structures can be understood in terms of structured bipolar argumentation frameworks. Models built on dialectical structures are thus related to models built on abstract, unipolar argumentation frameworks. Models with these argumentation frameworks have been explored by, among others, Borg et al. (2018), Butler, Pigozzi & Rouchier (2019), and Gabbriellini & Torroni (2014).

Owing to the different kind of argumentation frameworks they implement, these models are different to dialectical structure models in some respects. The epistemic environment is not composed of arguments as premise-conclusion structures, and agents can not build their belief systems in terms of individual propositions. Instead, agents perceive their epistemic environment to contain abstract arguments and relations between them. Since arguments are modelled

in the abstract, there is considerable flexibility in interpreting these epistemic environments. The objects could be interpreted as arguments, but also as reasons or hypotheses, objects that are quite unlike arguments. So write Borg et al. (2018, p. 288): "given the abstract nature of arguments, we interpret them as hypotheses which scientists investigate". Although these models do not implement arguments in terms of premise-conclusion structures, they are still models of argumentation in so far as they model the exchange of and relations between reasons. Figure 2.18 shows a simplified example for the epistemic environment implemented in Borg et al. (2018).

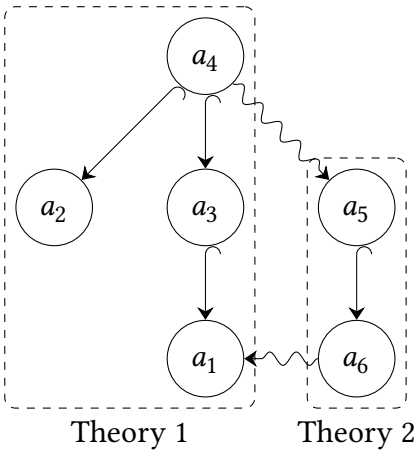


Figure 2.18: An argumentation landscape in Borg et al.'s model of abstract argumentation. Arguments are represented as nodes, the discovery relation with a hook arrow (↪), and the attack relation as a zigzag arrow (⇝).

Arguments in these abstract argument maps can be related to other arguments in two ways. The first such relation is the attack relation. In abstract argumentation frameworks, the attack relation determines whether arguments are "acceptable" (Dung 1995, p. 326), the underlying assumption being that agents would only accept arguments that are either not attacked at all or that can be defended through further arguments which attack the attackers. Like the arguments, this attack relation remains abstract. Abstract argumentation frameworks can not rely on the arguments' structure to explain why one argument would attack one another. These relations need to be determined by the model in a different way.

The discovery relation is the second relation that can hold between arguments in Borg et al.'s model. The existence of this second relation does not mean that this model implements bipolar argumentation frameworks. Unlike the support relation in bipolar frameworks, the discovery relation is more of a guide that tells agents how to traverse the argument map. It is not a relation between arguments in the same sense as the defeat and support relations in bipolar frameworks.

Any set of arguments in which no argument attacks any other is called a "conflict-free" set of arguments (Borg et al. 2018, p. 288). These conflict-free

sets are important to the model, because it uses conflict-free argument sets to model scientific theories. Not any conflict-free set of arguments can be a theory, though. Theories need to be connected sub-graphs in the abstract argument map that are connected through the discovery relation, and every argument in the map must belong to exactly one theory. Agents strive to discover the theory with the highest degree of defensibility (Borg et al. 2018, p. 289), or the theory with the most acceptable arguments.

At the start of a model run, agents in Borg et al. (2018) and similar models are aware of just a few arguments. For every argument they know, they are not necessarily aware of all attacks yet. Their beliefs at a particular moment in time are thus best described by the arguments and the attack relations that they are aware of at that point. Based on this belief system, the theory that would be (currently) preferred by the agent can be quickly determined.

Agents update these beliefs in two ways, by either discovering the attack relations that arguments have towards other arguments, or by learning about arguments and relations from others. Their preferred theory may change in this way, as further arguments and attacks can make a theory appear better defended. For example, in Figure 2.18, theory 1 would be chosen by fully aware agents as it has three acceptable arguments while its rival theory only has one. That does not mean that agents with a different view of this debate could not prefer theory 2. For example, consider an agent who is unaware of a_3 and a_4 as well as the attack $a_4 \rightsquigarrow a_5$. This agent would take theory 2 to have two acceptable arguments but theory 1 to only have one (a_2).

There are many questions that can be studied in this model. It is pretty clear that agents will eventually settle on the best theory if one theory is better defended than all others, but how long will agents need to converge? Will information learned from others accelerate this process? Is diversity in starting positions beneficial, and under which conditions?

2.5.2 Argument communication models

Argument communication models offer another way of implementing argumentation in agent-based models. In this group of models, arguments are modelled through the positive or negative influence they have on the agents' perspective of a proposition under discussion. Belief in this proposition is supported by arguments with positive valence, and disbelief is supported by arguments with negative valence. Agents accept or reject the proposition under discussion based on the arguments they are aware of, but they can only consider a limited amount of arguments in their decision making. These models are interesting because they show belief polarisation effects (Mäs & Flache 2013), and how agents could polarise under condition of epistemic rationality (Singer et al. 2019).

The name for this group of models is due to its best known representative, the *argument communication theory of belief polarisation* by Mäs & Flache (2013). In this section, I want to look at how this model and some of its relatives

(1) describe arguments and agents, and (2) which kind of argument exchange and belief updating events they implement.

Regarding the implementation of arguments, Mäs & Flache (2013, §1.3.2) describe a debate as being about an issue, to which arguments provide either a reason for acceptance or refutation. This leads to a two-tiered ontology consisting of issues and arguments. The pro- and con-relations are always directed from arguments towards the issue, but never between pairs of arguments or from the issue to one of the arguments. The positions of agents are represented by numeric values that reflect their stance towards the issue. New arguments can change the agent's stance by a numeric value based on the argument's relevance for the issue under discussion.

Banisch & Olbrich (2021) extend the model by Mäs & Flache (2013) and account for discussions with multiple issues. In their model, arguments can be related to more than one issue (Banisch & Olbrich 2021, §2.3): they may provide a reason in favour of one issue but against another one, reasons in favour of both, etc. Issues become related in this way, because an agent's acceptance of one argument can provide it with a pro-reason for one issue and, simultaneously, with a con-reason for another issue.

Figure 2.19 shows how argument communication and other models implement arguments as well as propositions and issues under discussion.

Argument communication models treat arguments uniformly without differences in argumentative properties. The models thus contribute to understanding the general influence of arguments on polarisation and other dynamical properties of belief, but leave open the influence that expressions of argument properties might have. For example, they do not differentiate the diverse intentions of arguments, which are captured as argumentation strategies in models of dialectical structures. Like abstract argumentation frameworks, argument communication models do not rely on a strict notion of argument as dialectical structure models do. This is shown by the fact that modellers in these other frameworks refer to the "arguments" in their models by other terms, such as "hypotheses" (Borg et al. 2018, p. 288) or "reasons" (Singer et al. 2019, p. 2244 et passim).

Argument communication models also limit the functions of arguments to providing reasons in favour or against an issue. Although these are clearly central functions of argumentation, not all arguments can be reduced to these roles: for example, arguments can shape debates by showing that the issues under discussion can be mutually accommodated, they can enlarge or reduce the scope of issues, etc. The effects on polarisation of these and other argumentative features are not investigated, and an argument exchange mechanism that resolves the inner workings of arguments seems necessary for this task.

Figure 2.19 also shows a graph representation of a model by Friedkin et al. (2016). Friedkin et al.'s model is a general model of opinion dynamics and, unlike the original argument communication model, not a specific model of polarisation. It is interesting when studying dialectical structure models though, because both encode logical constraints in belief dynamics: Friedkin et al. (2016) use a matrix C with elements $c_{ij} \in [0, 1]$ showing the logical constraint of

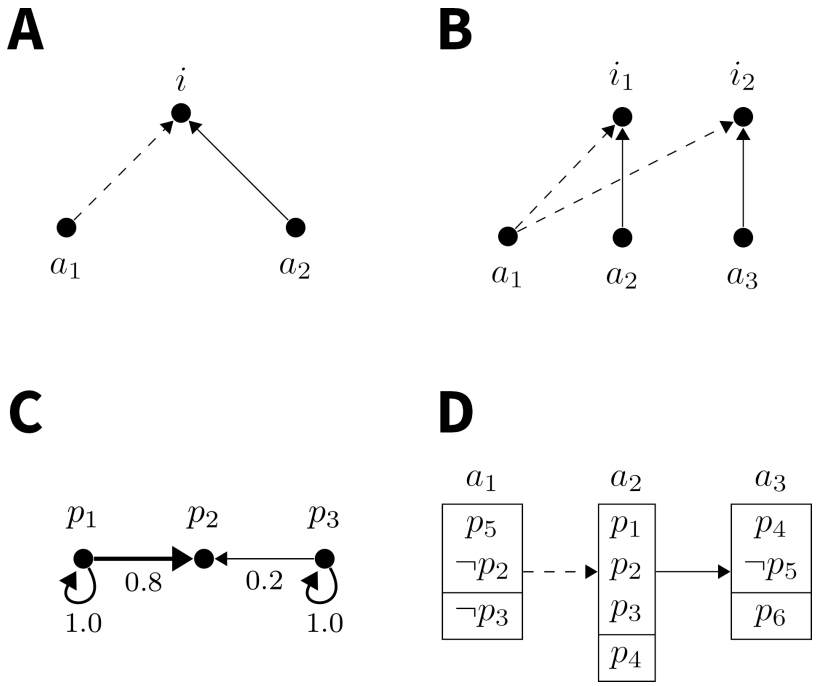


Figure 2.19: Illustration of selected models and their representations of the sentences under discussion, interpreted as directed graphs. (**A**) and (**B**) show Mäs & Flache's and Banisch & Olbrich's models, respectively. Those two distinguish issues i_j and arguments a_k. Dashed edges show arguments supporting a negative stance towards the issue (contra arguments), and solid edges support a positive stance (pro arguments). In (**B**), more than one issue is considered in a debate, but not every argument needs to contribute to all issues. (**C**) illustrates Friedkin et al.'s model, in which nodes represent propositions and edges show the weight of doxastic implication: in this example, an agent's belief in p_1 implies belief in p_2 to a degree of 0.8, and an agent's disbelief in p_1 would imply disbelief to the same weight, but its belief in p_1 is not constrained by belief in another proposition. An example for debates in the dialectical structure model is in (**D**), where nodes are arguments consisting of premises and conclusions, and edges are either of the support or defeat relation. (Figure from Kopecky 2022)

sentence i on sentence j, while dialectical structure models use Boolean formulas to represent logical constraints. In Friedkin et al.'s model, the effects from the logical constraints in the matrix C compete with influences from a social network in forming agents' belief systems. The matrix C is stochastic, which requires that the row sums equal 1. In the graph representation, this translates to all constraints on a proposition adding up to 1 in their weight, including reflexive constraints.

Beside their modelling of the epistemic environment, a second important difference among agent-based debate models is how they have agents behave towards each other and how agents update their belief systems. Mäs & Flache's and Banisch & Olbrich's are social influence models (see Flache et al. (2017) for a typology). In these models, the opinions that agents move to in an updating event are aggregates of the distances between the updating agent and the agents that influence it. In contrast to this updating mechanism, agents in dialectical structure models are only indirectly influenced by the belief systems of other agents, through the arguments those agents introduce. This difference is representative of the general distinction between social and informational influence in agent-based modelling and beyond (Burnstein & Vinokur 1977). In models of informational influence such as dialectical structure models, agents decide how to change their opinion based on newly introduced arguments and maximal opinion continuity, and do not rely on particular neighbours for their updating.

Argument communication models do not only describe *how* agents' belief systems are influenced by others, but also *which* agents can exert this influence. Mäs & Flache's and Banisch & Olbrich's models only yield polarised outcomes when they rely on homophily to determine which agents are partnered up in an argument exchange event (see Figure 7 in Banisch & Olbrich 2021 and Figure 3 in Mäs & Flache 2013 for the effects of homophily in their models). In the context of these models, homophily basically means that agents are more likely to communicate the more alike they are. While this influence is a well-established phenomenon in the communication of humans and an interesting factor in its own right (McPherson, Smith-Lovin & Cook 2001), its influence may limit the insights to be gathered about *arguments* as drivers of polarisation. Homophily, after all, is not a property of arguments, but of the agents. Dialectical structure models are different in that respect: they exhibit rising polarisation even as all agents communicate in a continuous debate forum, with all of them having equal probability to select each other as communication partners.

Singer et al. (2019) present a model that can be understood as a variant of the argument communication models by Mäs & Flache (2013) and Banisch & Olbrich (2021). While these latter models follow sociological research interests, Singer et al. have philosophical intentions. A comparison between these models elucidates the different goals and constraints of computational modelling in sociology and philosophy. This points us to the specific and substantial contribution that philosophical models can offer. Like Mäs & Flache, Singer et al. determine their agents' beliefs through aggregating a set of currently possessed arguments. In Singer et al., agents hold beliefs in $[-n, n]$ that reflect the

strength of possessed reasons, while beliefs are normalised to [0, 1] by Mäs & Flache. In both models, agents can remember a limited number of reasons (6 in Mäs & Flache, 7 in Singer et al.). But there are noteworthy differences in how agents communicate their beliefs with others and manage their limited memory. While the communication in Singer et al. (2019) is public, 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. They forget the reason they held for the longest time in Mäs & Flache (2013), irrespective how well it supports their current opinion. In Singer et al. (2019), agents forget the weakest argument that contradicts their current opinion on the issue under discussion.

Although the models share many formal similarities, it would be difficult to make the case for epistemically rational polarisation for Mäs & Flache (2013). Public and unbiased communication as well as coherence-minded updating are important features in Singer et al.'s model that make their claim of epistemically rational belief polarisation more plausible. This coherence-minded updating process is shown in Figure 2.20. At time t_1, agent a_3 shares an argument, or "reason", of strength 7 with all other agents. As agent a_1 has a positive opinion at time t_1, it forgets about a reason with negative valence, starting with the weakest such reason. Agent a_2 likewise forgets the weakest reason that contradicts its current view. Note though that a_2 only possesses reasons with negative valence at t_1. While a_1 strengthens its previous opinion through this exchange, a_2 moderates it.

A noteworthy assumption in Singer et al.'s model is that arguments have objective strength. Following this assumption, arguments are strong or weak independently of how the agents evaluate them. For example, in Figure 2.20, the reason shared by a_3 has the same strength 7 for the sender and for both receivers. An alternative assumption could be that agents evaluate arguments differently depending on their prior stance on the issue. Mäs & Flache (2013) make a much weaker if still similar assumption, where arguments only have objective veracity in ± 1. It would be interesting to observe effects of agent-relative argument strength in these models.

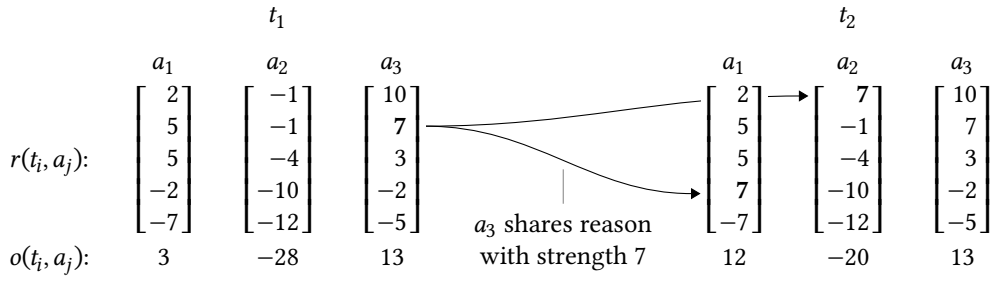


Figure 2.20: Coherence-minded updating in Singer et al. (2019).

Does it matter for Mäs & Flache (2013) that their model does not allow for epistemically rational polarisation? Not really. They pursue a different research question. Mäs & Flache are looking for contributors to bi-polarisation, irrespective of whether it is brought about rationally. In a bi-polarised population, positions on the end 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.

2.5.3 Bayesian epistemology

Bayesian approaches to beliefs in humans can be easily distinguished through a few core assumptions: that the beliefs of agents can be expressed (a) in degrees of belief, that (b) these degrees satisfy the axioms of probability theory and that (c) updating one's degree of belief follows a rule of conditionalisation such as given by Bayes' theorem or Jeffrey conditionalisation (e.g., Titelbaum 2022). In agent-based models, this translates to multiple agents having degrees of belief that satisfy the probability axioms and that are updated through conditionalisation in each round. Consider that an agent's belief in a hypothesis h at the initial step i was given by $p_i(h)$. The new evidence e on which the agent conditionalises its prior beliefs may consist in social or informational influence. The updated belief $p_u(h)$ after conditionalising on this evidence is then given as $p_i(h|e)$.

This representation of beliefs differs from the models discussed so far. From a Bayesian perspective, agents' beliefs are modelled as probabilities in $[0, 1]$ (Dorst 2023; O'Connor & Weatherall 2018; Olsson 2013; Pallavicini, Hallsson & Kappel 2021). The argument communication model from Singer et al. (2019) uses a range of real numbers $[-n, n]$ instead, where n is determined by number and weight of reasons. And in dialectical structure models, beliefs are multidimensional, categorical mappings from the propositions under discussion to truth values.

There are different ways to spell out the conditionalisation process necessary for updated beliefs. A common rule for conditionalisation is given by Bayes' theorem:

$$p_u(h) = p_i(h|e) = \frac{p_i(e|h)p_i(h)}{p_i(e)}$$

Against Bayesian methods, it is sometimes argued that it makes false assumptions, either because rational agents like humans did not have beliefs that could be described as credences in $[0, 1]$ or that we humans could never really determine these credences. In the context of agent-based modelling I would like to reject this criticism. Every modelling approach makes assumptions about beliefs and the rational behaviour of agents. All of these assumptions risk conflating important details and they all may fail at some point. These abstractions are tolerable as long as the framework can provide a useful and reliable approach for the theoretical study of epistemic rationality.

When conditionalisation is based on the opinions of others mediated through the trust one agent has in its peers, the modelling approach can be called "models of source reliability" (Merdes, von Sydow & Hahn 2021). Source reliability is a similar, but different, factor compared to homophily in argument communication models. In models that rely on homophily, agents are more likely to influence the beliefs of others if their beliefs are more similar. In models of source reliability, agents instead condition on the beliefs of others based on their trust – which may or may not coincide with them holding similar beliefs.

There are quite a few models that use Bayesian conditionalisation to generate belief polarisation (O'Connor & Weatherall 2018; Olsson 2013; Pallavicini, Hallsson & Kappel 2021), though none of these claim epistemic rationality for this polarisation process. And it is easy to see why: trust is really a limiting factor to an agent's ability to engage with its epistemic surroundings. Trusting some too much and others too little can lead an agent to become unresponsive to evidence that would have changed its mind. A noteworthy exception to this rule is Dorst (2023), who claims epistemic rationality for belief polarisation in a Bayesian model. The ambiguity of evidence, Dorst says, can lead rational agents to pursue different paths and end up in very different locations. Dorst's Bayesian model of rational polarisation and the dialectical structure model from Chapter 4 thus give mutual, inter-methodological support for the possibility claim of epistemically rational belief polarisation.

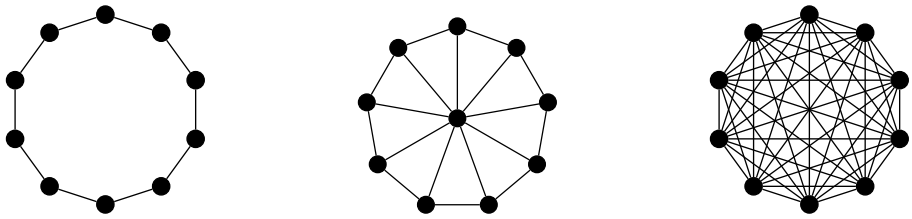
One issue for some Bayesian models is that they only track belief dynamics in one proposition, and it is a plausible assumption that some epistemic mechanisms will change if more propositions are involved. But Bayesian epistemology also offers a method to achieve reasoning about multiple propositions in terms of Bayesian nets (Bovens & Hartmann 2003, §3.5). There is at least one application of this technique in agent-based modelling (Grim et al. 2022).

2.5.4 Network epistemology

Zollman (2007) wondered how the communication among epistemic agents should be organised so that they converge quickly and arrive reliably at the truth. Not surprisingly, agents converge to agreement quicker if they are better connected. But networks with less connectivity provide a more reliable communication structure for agents to settle on the correct answer to a problem. There seems to be a tension between getting things done quickly and getting them right.

In network epistemology models, a group of agents inquires about a received theory and its alternative. Agents hold beliefs in $[0, 1]$ about the quality of the alternative theory. Agents that hold a belief in $[0, 0.5)$ will stick to the received theory, and agents that have a belief in $[0.5, 1]$ will propose the alternative. In the simulated world the alternative theory is better to a slight degree. But of course Zollman does not tell his agents about this. Initially, agents' beliefs are sampled from a normal distribution in $[0, 1]$. Each round, agents test their

currently preferred theory and receive a signal based on the objective, but unknown, success rate of this theory. They also include information obtained by their peers, where a "peer" is any agent that is connected to them in a communication network. The communication network is antecedently determined and held fixed throughout a model run. The model thus combines informational and social influence. Zollman (2007, p. 579) gives three archetypical communication networks (see Figure 2.21), although these are only exemplars for the density, or connectedness, of a social network. Zollman did not rely on the specific structure of these archetypical networks and confirmed the effect for all possible networks of six agents.



(a) Circle (b) Wheel (c) Complete

Figure 2.21: Three prototypical communication networks as given by Zollman (2007, p. 579)

Zollman (2010) found that well-connected, or "dense" networks, such as a complete network (Figure 2.21c) are very conducive to quick convergence to consensus. However, that consensus may be a premature consensus, because it can make agents blind to evidence that is not exhaustively explored. On the other hand, sparse networks such as the circle (Figure 2.21a) and, to a lesser extent, the wheel (Figure 2.21b) promote slower, but more reliable convergence to the objective credence. The reason for this, Zollman (2010) argues, is that these latter networks allow for a longer upholding of diverse approaches to a problem, which guarantees that at all evidence is properly explored. Maintaining diversity for some time – but not indefinitely – thus appears to be epistemically beneficial (Zollman 2010).

It should be noted that Zollman's results do not seem to hold in all ranges of the parameter space (Rosenstock, Bruner & O'Connor 2017), and that he did not understood diversity in terms of a diversity index such as those discussed in Section 2.3. On the other hand, it does not seem obvious that these findings could not be replicable if a received diversity index was applied, and it is not required for an effect to hold in all possible conditions to be a real effect.

Simulations on this model illustrate the benefits of what Zollman (2010) calls "transient diversity". Allow a group to pursue diverse approaches and converge only after sufficient evidence exploration, and you protect against the falsification of initially popular hypotheses. The results about a high inconsistency prevalence in groups with high opinion diversity, obtained in Chapter 5 in this thesis, are complementary to these results. In the scenario envisioned in Chapter 5, groups of agents evaluate evidence under tight time constraints. In

this exceptional scenario where no new evidence can be gathered and communicated to others, diversity does not seem to be beneficial. But these scenarios preclude all evidence accumulation and exchange that Zollman allows his agents to perform. A combined evaluation of these two models suggests that the benefits of diversity can only be reaped after considerable investigation and exchange.

2.5.5 Epistemic landscapes

Like the network epistemology approach, epistemic landscapes have been applied to study the effects of diversity in agent-based models. The most popular application of this approach is due to Hong & Page (2004) (with critical reception in Thompson 2014 and Grim et al. 2019), but it is pursued broadly such as by Weisberg & Muldoon (2009) or Pöyhönen (2017).

An epistemic landscape can be described through a function $V : \mathbb{N} \to \mathbb{R}$ that maps a finite number of epistemic profiles to real-valued epistemic payoffs (Hong & Page 2004, p. 16386). An epistemic profile can be interpreted as an opinion, a research programme, a hypothesis, etc. A payoff is the success associated with the profile. For example, the epistemic payoff associated with an opinion could be the likelihood that this opinion turns out to be true. Agents are equipped with an initial profile and seek to optimise their payoff by traversing the landscape with limited sight. There is a straightforward and accessible visual representation of epistemic landscapes, shown in Figure 2.22. Note though that epistemic landscapes are rings and that the left-most point in the plane is adjacent to the right extreme.

In Figure 2.22, the optimum shown in the displayed section is the position to the right of agent *a*, so it is plausible to assume that the optimal solution would lie in *a*'s field of vision. Agent *b* is in a more difficult position – not only is the optimal solution far away, there a local optima to the right of *b* which could lure *b* away from the optima to the left (assuming that the counter-clockwise move towards *a* would be the shortest route for *b* to approach the global optimum).

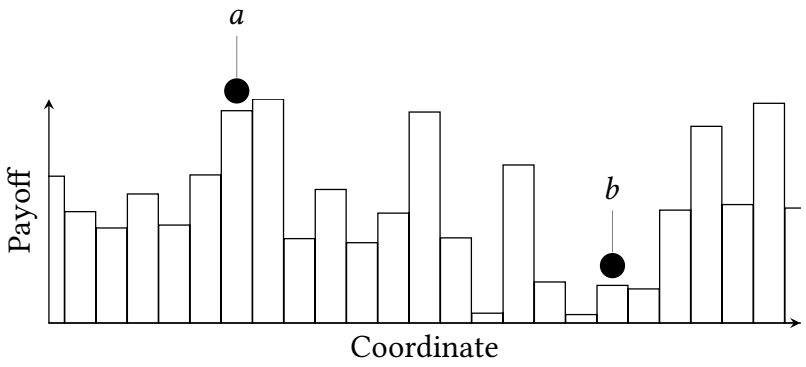


Figure 2.22: A two-dimensional epistemic landscape

Now imagine that there are more agents than just *a* and *b*. Assuming that none of the agents are initially aware of the global optimum, how should a sample of agents be distributed along the epistemic landscape to maximise their chance of eventually finding the optimum? A plausible hypothesis would be that it should be the agents that are most competent in finding a high payoff, but Hong & Page (2004) find that a random sample in many cases outperforms a sample of most competent but imperfect agents. This result is a tricky one. It does not seem to hold in the complete parameter spectrum, and is particularly sensitive to the "ruggedness" of a landscape (Grim et al. 2019).

Like in models of networks epistemology, the epistemic landscape models of Hong & Page (2004) and Pöyhönen (2017) yield results that suggest an epistemic benefit of diversity in epistemically relevant factors. This finding seems to be in conflict with the findings from Chapter 5 of this thesis, which indicates that some decision tasks are more difficult for very diverse groups. But like in Zollman (2010), the benefit to diversity in epistemic landscape models is inextricably linked to time. The benefits of diversity can not be reaped early in these models. Despite issues under exceptional circumstances, such as limited time constraints, many results from computational social epistemology suggest diversity benefits under normal circumstances that allow for sufficient evidence gathering and communication.

A critical note is due here regarding Hong & Page's way of determining whether a sample is diverse. The index of diversity they use is unusual and could be expressed as the average pairwise normalised Hamming distance (Hong & Page 2004, p. 16386). Based on my discussion of measures from Section 2.3, this would be much more like a measure of disagreement. Hong & Page certainly studied *differences* in epistemic profiles, but I am not convinced that they studied *diversity* in the sense of an index of diversity. A similar worry is expressed by Thompson (2014, pp. 1027–1028), who is generally critical of Hong & Page's findings. Some of her criticisms have in turn been critically received in the literature (Singer 2019) and a definitive verdict on diversity and epistemic group performance remains to be written.

2.5.6 Lehrer & Wagner's "rational consensus"

In the late 1970s, Lehrer & Wagner (1981) performed some of the earliest application of computer simulation in philosophy and laid the foundations for a model that continues to be studied (e.g., Hartmann, Martini & Sprenger 2009). Their model shows that convergence to an agreed position from initially disagreeing beliefs is possible through continuous weighted averaging. This averaging is successful when all members of the group can assign an initial belief in $[0, 1]$ to the issue under discussion and can also assign positive, non-zero weights in $(0, 1]$ to the beliefs of others. The model is not guaranteed to converge if some agents assign zero weight to the belief of other agents (Lehrer & Wagner 1981, pp. 26–27), and convergence may take some time to unfold.

This model describes a reliable process in which agents that are competent to judge an issue and the opinions of others could settle their disagreement.

Lehrer & Wagner assign strong normative force to these results. Lehrer (1976, p. 327) called groups that meet the requirements of the model but did not converge towards consensus "demonstrably irrational". How many groups really meet these requirements and whether the time to convergence is suitable to practical purposes is, of course, a different matter.

An example of this model can be easily constructed. Consider the following matrix W of weights that agents assign to their peers and the initial credences in P. The matrix W is stochastic with row sums of 1, where the first row collects the weights that the first agent assigns to all other agents – including the weight the agents assigns to itself in the first column.

$$W = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}, P = \begin{bmatrix} 0.5 \\ 0.9 \end{bmatrix}$$

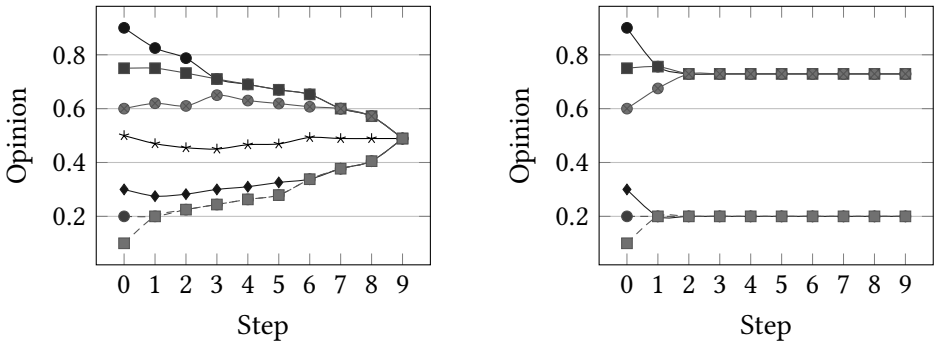
In the first iteration, this scenario would yield beliefs of 0.58 and 0.78 for the first and second agent, respectively. This result is obtained by the product WP. The second iteration can be obtained through $W(WP)$, the third through $W(W(WP))$, etc. In general, the nth iteration of Lehrer & Wagner is given by W^nP. As $n \to \infty$, the agents in the example above converge to $[0.66, 0.66]$. An approximation of this consensus is reached at $n = 6$ ($[0.6575, 0.66375]$), though this procedure might take much longer in other scenarios.

2.5.7 Bounded confidence and low resolution modelling

In the second half of the 20th century, sociologists, philosophers, and mathematicians began to look for factors that could determine under which conditions belief dynamics in artificial societies would converge to consensus, and under which conditions would yield polarisation. Hegselmann & Krause (2002) found that such a determining factor could be a bounded confidence parameter in belief averaging. Consider that you would hold a belief in $[0, 1]$ and that you would update it by averaging the opinions of yourself and your neighbours, but that you would not consider everyone in the population a neighbour. Whether the beliefs of someone else would influence your opinion would rather depend on a parameter of opinion distance. The larger this parameter, the larger is the amount of people that you would eventually find yourself in agreement with.

The outcome of belief averaging with bounded confidence also depends on the initial distribution of beliefs. If the beliefs are spread evenly throughout the space of available opinions, a population can achieve consensus through a relatively low bounded confidence parameter. In Figure 2.23a, a bounded confidence parameter $\varepsilon = 0.2$ is sufficient to yield consensus. A consensus is impossible to achieve, however, with a $\varepsilon = 0.2$ in a population that lacks the agent marked by a star. Without the star, the belief dynamics of the population in Figure 2.23b yield stable belief polarisation.

In reflections on their model, Hegselmann & Krause (2009) praise the "low resolution" modelling approach enabled through the bounded confidence parameter. Their approach indeed has a lot to recommend itself. The bounded confidence parameter and a minimal representation of belief systems as real numbers allow for easy inspection, replication, and understanding of the



(a) This population of seven agents converge to population-wide agreement through consecutive averaging.

(b) If the star agent is removed from the former population of seven, this population of six is unable to reach agreement if the bounded confidence parameter is not increased.

Figure 2.23: Belief dynamics obtained from averaging with bounded confidence of $\varepsilon = 0.2$. Each agent has an opinion in $[0, 1]$ represented by a symbol in the plot. At each step, every agent looks for beliefs with a maximum distance of ε to its own and forms its new belief based on the average of beliefs within this radius.

model. It offers a simple mechanism for the complex question of why belief dynamics in humans sometimes converge to unanimity and sometimes to disunity. But while this approach serves their research question well, it does not serve others at all. Agents in this model can hardly claim to meet epistemic rationality constraints. They disregard far-removed beliefs only based on the fact that they are far removed – from themselves. This thesis can be read as a study of more demanding phenomena through more complex computational models. And the results allow for an optimistic outlook: by increasing the resolution of computational models we can understand more complex phenomena.

Chapter 3

Arguments as drivers of issue polarisation

3.1 Introduction

Two recent agent-based models of polarisation (Banisch & Olbrich 2021; Mäs & Flache 2013) rely on argument communication as a driver of polarisation. These studies underpin the hypothesis that arguments and their properties could play a role in polarisation dynamics, alongside diverse other candidate causes, such as lacking exposure to other views (Mutz 2002) or the effects of one-to-many communication in online networks (Keijzer, Mäs & Flache 2018).

Studying polarisation in agent-based debate models can profit our understanding of polarisation in humans in two ways. First, agent-based models can help in formulating hypotheses about which factors contribute to polarisation in humans, which then can be tested in experimental studies on human deliberation. Secondly, since agents in debate models exhibit an idealised version of rational deliberative behaviour, debate models can help us understand which extent and kinds of polarisation we will have to expect even in cases in which human behaviour approximates this kind of idealised rationality.

This chapter further elucidates the hypothesis that arguments can drive polarisation: argumentation on its own can be a driver of polarisation in populations of artificial agents even when communication between agents is not governed by social influence. In models of social influence, an agent's updating is shaped by what other agents think, in particular its close neighbours. This chapter presents results from a type 1a dialectical structure model (see Section 2.4.1). In this model, agents only listen to what the arguments say when updating, and they influence the updating of others by introducing logical constraints on which opinions they can choose.

Simulations on this model exhibit rising polarisation after argument introduction in three standard polarisation measures. But not all arguments affect polarisation equally: different strategies of selecting premises and conclusions for an argument can influence the obtained polarisation values. The measures show noteworthy differences depending on the argumentation strategies employed in premise selection. When agents take into account the opinion of a communication partner as premises ("allocentrism"), their arguments have an increased chance of facilitating low values of polarisation. Arguments that take

off from premises the agents themselves hold ("egocentrism") have a lower chance of inducing low polarisation. Additionally, when models are initialised with perfect bi-polarisation between groups, the allocentric strategies are able to de-polarise the debate towards medium levels of polarisation, while the egocentric strategies appear unable to recover from bi-polarisation.

This chapter is organised as follows. Section 3.2 briefly recaps the theoretical foundation of the model from Chapter 2. The evaluation of simulation experiments that track measures of issue polarisation on this model are discussed in Section 3.3, and a robustness analysis in Section 3.4. I discuss the results and the limitations under which they are gathered in Section 3.5.

3.2 A brief description of the model in use

This chapter shows the results from simulations on a type 1a dialectical structure model. In this model, agents engage in a debate with 20 atomic sentence variables. This sentence pool remains unchanged throughout all model runs. Agents hold binary beliefs towards all sentences under discussion. Unless otherwise noted, their beliefs are initially drawn at random. The argument maps evolve through the introduction of arguments by agents. Argument maps are empty initially and grow randomly, as illustrated in Figure 2.7. Agents are drawn randomly to introduce arguments according to one of five argumentation strategies, those introduced in Section 2.4.1.2, page 54. All agents use the same argumentation strategy in each model run, though they may introduce different kinds of arguments through the any strategy. The arguments agents introduce have two or three premises and never introduce satisfiability violations. After each argument introduction, agents verify that their beliefs are still consistent with the debate and update to the nearest neighbour in case they are not.

The purpose of the simulation experiments reported below is whether and under which conditions arguments can drive belief polarisation in artificial agents.

A full description of this model can be found in Chapter 2. In addition, a description within the ODD protocol is given in Appendix A. The supplementary materials mentioned at the end of this chapter (Section 3.6) contain Jupyter notebooks for the replication of the results presented here.

3.3 Experimental results

3.3.1 Experimental design

Since argument introduction and position updating include elements of random choice, it is unavoidable to study dialectical structure models in simulation experiments with many iterations. In this section, I present the results of seven experimental settings. The main experiment has a population of 50 agents and 20 atomic sentence variables, resulting in 40 sentences available as

premises and conclusions. In a robustness analysis, I also study the model in conditions of (1) 20 agents and 20 sentence variables, (2) 20 agents and 10 sentence variables, (3) 10 agents and 5 sentence variables (resulting in 10 available sentences), (4) 50 agents and 20 sentence variables, but an argument length of 2–4 premises instead of 2–3, (5) with initial positions in perfect bi-polarisation and (6) with a clustering on a subset of key issues. This robustness analysis does not only confirm the results from the main experiment, but shows that polarisation effects can be amplified by contraction of the population and in particular the sentence pool.

There are a total of 5,000 experiments in each setting, 1,000 for each argumentation strategy. Apart from varying population size, sentence pool, and, in one case, length of arguments, all experiments have the same set-up: all five argumentation strategies are compared in each experiment and the experiment always runs until either a density of $D = 0.8$ is reached or an argument introduction fails, whichever occurs first. In the main experiment, all debates end due to the density condition. Termination there occurs on average after 110 turns for convert model runs, 91 in undercut, 131 in fortify, attack models take 126 on average, and models with the any strategy 92. All in all, the data for the main experiment consist of 548,958 debate stages. I evaluate the model runs by applying three polarisation measures, all of which are adapted from the definitions in Bramson et al. (2016): dispersion, group divergence and group consensus. All measures return values in the $[0, 1]$ range.

The observed values for issue polarisation are lower in this model than in other studies, and only a low percentage of simulations end in clear-cut bi-polarisation (which happens frequently in the social influence and Bayesian models discussed above). But it is important to keep in mind that this model only accounts for the influence of arguments and argumentation strategies on polarisation, and drivers such as homophily as well as other properties of agents are ignored.

3.3.2 Dispersion, understood as standard deviation

Dispersion tracks an intuitive idea of how agents and their belief systems can polarise by measuring how agents deviate from a population-wide mean. If these spread out evenly or cluster around one pole, dispersion will be low, but clustering around an increasing number of poles will lead to increased dispersion.

I introduced a measure of dispersion, adjusted to dialectical structure models, in Chapter 2, Definition 8 (page 46). It does not require an antecedent opinion clustering. The measure can be directly applied to the beliefs held in the population.

Figure 3.1 shows the development of dispersion depending on argumentation strategy and plotted against density in the main experiment. It shows that the introduction of arguments generally increases polarisation. The argumentation strategies differ in their contribution to polarisation, which is

comparatively high in the attack and comparatively low in the convert strategy. Agents with the any strategy show a slightly higher rate of polarisation in lower density, but end up less polarised than the attack and fortify model runs. This seems to indicate that the different effects of the strategies seem to balance each other out when triggered in alternation.

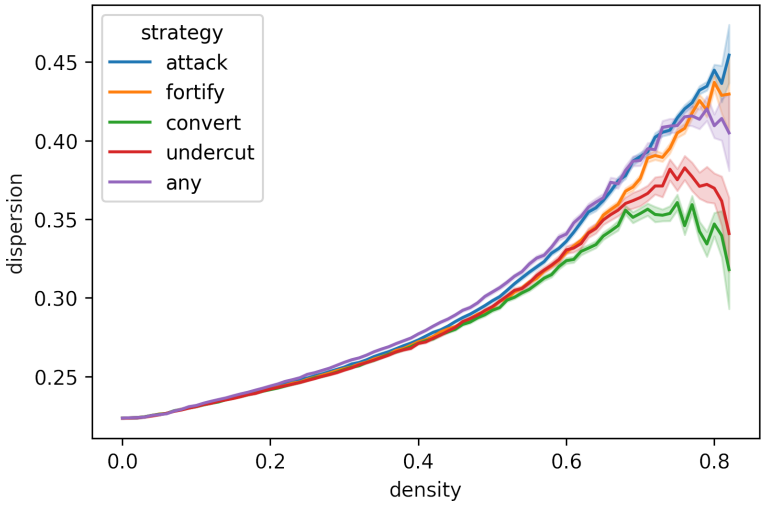


Figure 3.1: Mean dispersion plotted against density and depending on argumentation strategy in the main experiment. Shaded areas show confidence intervals based on statistical bootstrapping. *Note*: The original values reported in Kopecky (2022) were performed with a calculation error. Please see https://www.jasss.org/26/2/3.html for the correction notice in the journal.

When simulations reach densities of around 0.8, the model runs are about to terminate after each agent has introduced 1–2 arguments on average, and the effects of argumentation will be most visible in this period. At densities of around 0.8, the mean values for attack simulation runs are higher than in other strategies, particularly for convert and undercut. The latter more often reach lower dispersion values than the other strategies, and some of the simulation runs have dispersion values comparable to their initial values, going against the general tendency.

3.3.3 Group-based measures

Dispersion measures pairs of agents uniformly and is ignorant as to whether they are members of different communities, or groups. Group-based measures (as defined by Bramson et al. (2016, pp. 89–93)) rely on the community structure, or clustering, of the population and treat distances between members of the same group different to members of different groups. Groups can be determined either endogenously or exogenously (Bramson et al. 2016, pp. 87–88). An endogenous definition works on the structure of the population alone, such as in community structuring algorithms. Exogenous definitions require the addition of pre-defined criteria to partition the population into groups.

In Section 2.3.2, I discuss the application of two state-of-the-art clustering algorithms to dialectical structure models, Leiden (Traag, Waltman & van Eck 2019) and affinity propagation (Frey & Dueck 2007). I used the implementations from `python-igraph` version 0.9.1 (Csárdi & Nepusz 2006) and `scikit-learn` version 0.24.1 (Pedregosa et al. 2011), respectively. I mostly rely on Leiden clusterings in the results reported below, while affinity propagation was used to compare and legitimise the Leiden clusterings.

Recall from Section 2.3.2 that the input to each clustering algorithm is an adjacency matrix of the population of agents, where the values are based on the Hamming distance between the agents' positions, normalised by the number of atomic sentences (i.e., $\mathrm{HD}(x, y)/20$ for agents x, y in the main experiment). For Leiden and affinity propagation, these distances were transformed by $\exp(-4x)$, and resulting values below 0.2 filtered out. The clustering success rate in Leiden is consistently above 90%. Transformation and filtering also improve the convergence rate for affinity propagation, which then is around 80–90%. Both algorithms output non-overlapping clusters and are deterministic, i.e. they output the same clustering for the same input every time they are run.

Since there are no previous reports of applying Leiden and affinity propagation on dialectical structure models, I used the adjusted Rand index (ARI, Hubert & Arabie 1985) to ensure the reliability of the clusterings. The ARI compares two clusterings by looking into how many agents are clustered into the same group in both clusterings, and how many are clustered into different groups. For the present purpose, I apply the ARI to count how many pairs of agents that are clustered into the same community in one debate stage are also members of the same community in the following debate stage, thus measuring in how far an argument introduction changes the clustering. A low ARI indicates that many agents have been clustered differently compared to the previous debate stage, while a higher ARI shows that more agents are in the same cluster as before, thus implying a lower mobility of agents and less force of arguments to influence the composition of groups. Then, the goal is to have a somewhat high, but not too high mean ARI value that can confirm the intuitively plausible expectation that the majority of agents remain in their group in most debate stages. The model should also allow for some fluctuation in the ARI, because some argument introductions have little if any effect on the debate, while others convince many to change their views. In the evaluation of the model, the ARI between pairs of adjacent debate stages took a median value of about 0.7, depending on the argumentation strategy (see Figure 2.14). The observed values indicate that clusterings based on the model are stable enough to simulate intuitively plausible opinion dynamics, and are thus reliable.

3.3.3.1 Group divergence

An interesting question to ask about a population's community structure is how far apart its groups are, or what their degree of divergence is. Recall

from Section 2.3.3 (Definition 6) that group divergence indicates this divergence as the amplitude of difference between in-group agreement compared to the strength of disagreement among agents that are in different groups.

Before going into the analysis on averaged values from large amounts of simulations, let me present the clustering analysis and resulting divergences in single runs of the model. Figure 3.2 looks at two single runs, one in which agents only use the attack argumentation strategy and one in which they only use convert. The figure shows how these populations of agents move into different states of polarisation. These can be considered as typical evolutions for the attack and convert strategies. While all strategies have a non-zero chance of ending in low or moderately polarised states, low polarisation is much more likely in convert, and to a lesser extent in undercut debates, and moderate polarisation is most likely in attack, and somewhat more likely in the fortify debates. So the evolution shown in Figure 3.2 for attack could have materialised with a different strategy – but it is more likely for an attack debate to behave in this way.

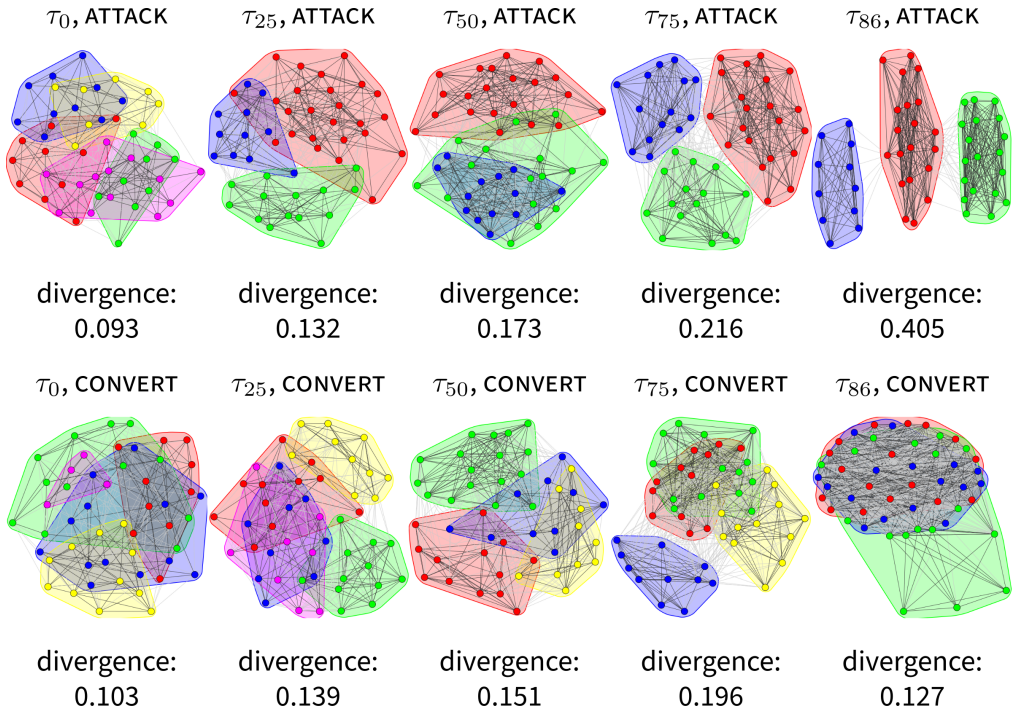


Figure 3.2: Development of clustering and group divergence for two sample debates through debate stages τ_i. The upper sample (attack strategy) shows tri-polarisation and monotonously increasing divergence values. The lower example (convert strategy) shows lower polarisation and signs of convergence. Figure from Kopecky (2022).

Probably the most interesting feature of the attack series in Figure 3.2 is tri-polarisation in the last debate stage – especially when it is contrasted with convergence in the last debate stage of the convert series. There are two other differences. First, notice how the divergence is steadily increasing in the attack run, but developing non-monotonously in the convert run. This seems to

show how the convert strategy is able to recover from increasing polarisation. Secondly, while both runs start with a high number of groups (5), the attack strategy very quickly reduces to only 3 groups, while the convert strategy is able to maintain its diversity until τ_{25}, and is able to uphold 4 groups until at least τ_{75}. This ability to maintain a higher diversity could be interpreted as contributing to the lower values observed in convert simulations.

From the main experiment, the overall results for divergence depending on the two clustering algorithms are shown in Figure 3.3. As in the results for dispersion, these show how the introduction of arguments contributes to polarisation in general. More particularly, the attack strategy shows the highest polarisation values, while values particularly in undercut and convert seem to be more frequent in less polarised states. Values for the any strategy also lie between those of the four basic strategies, confirming the observation from the dispersion measurement. All together, this confirms the overall pattern from the dispersion values.

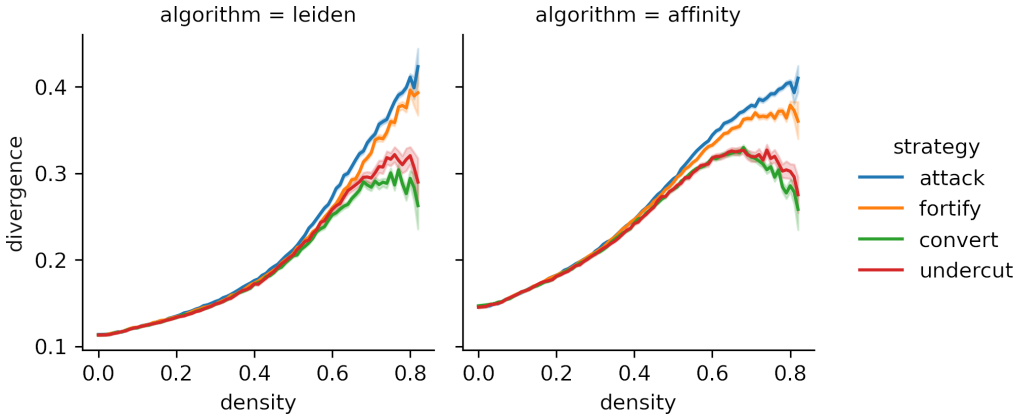


Figure 3.3: Mean divergence plotted against density, depending on argumentation strategies and clustering algorithms. Shaded areas indicate confidence intervals in terms of statistical bootstrapping. Figure from Kopecky (2022).

Figure 3.4 takes a more concentrated look at the divergence data by showing the divergence distribution for the simulation runs as they reach a density of around 0.8. The panes compare the main experiment with two robustness analyses, and they show a noteworthy difference among argumentation strategies: convert and undercut model runs reach low levels of divergence much more often, and they have smoother distributions, whereas fortify and particularly attack model runs are single-peaked with a considerably lower chance of ending in low polarisation, an effect that remains stable in the robustness analyses (to be further discussed in the dedicated section below).

The group divergence analysed through Leiden clusterings can be further quantified. 16% of simulation runs in the convert strategy have a group divergence of less than 0.2 – which is a very low increase, if any at all, from the start of the debate. For the attack strategy, only 0.8% of simulation runs reach

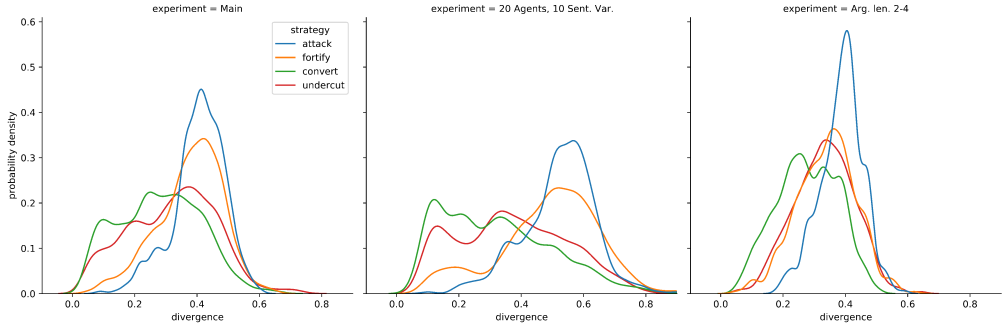


Figure 3.4: Distribution of divergence values based on Leiden clusterings at densities of around 0.8 for individual simulation runs in three different experiments. The area under each graph adds up to 100% of the divergence values, and the individual graphs show how the divergence values are distributed. For example, it shows that the convert strategy has a much higher proportion of model runs with a divergence of less than 0.2 at density 0.8 than the attack strategy, and that the biggest proportion of convert runs (about 20%) in the main experiment has a divergence of around 0.3, while the attack strategy peaks at around 0.5 with a proportion of more than 30%. Figure from Kopecky (2022).

this low level of polarisation. But about 43% of simulation runs with the attack strategy show moderate polarisation of at least 0.4, while only about 19% of convert debates do so. Fortify and undercut strategies are somewhere in between, with the fortify debates showing more tendency for medium polarisation and the undercut showing at least some chance of lower polarisation. The divergence mean for all data points at a density of around 0.8 is 0.29 for convert, but 0.41 for attack (see Table 3.1).

Table 3.1 provides a possible explanation for the lower mean polarisation in convert and undercut: it might be due to their increased ability to reach states of very low polarisation more often (the lowest 10% of both strategies have a mean of 0.08, compared to 0.24 in attack). However, the strategies seem to differ less in their chance to reach higher values of polarisation, as the highest 10% of simulation runs have lower variation.

Qualifying the result that arguments drive polarisation on their own, two argumentation strategies, convert and undercut, have a higher chance to end in states of low polarisation, while the other two, fortify and attack, tend to drive moderate levels of polarisation. These tendencies are noteworthy because they run parallel to another distinction: fortify and attack are very much egocentric argumentation strategies insofar as they select premises from the source position. Convert and undercut are allocentric strategies by the same standard: the source devises an argument with premises that the target accepts. So it seems that, in agent-based debate models, egocentric premise selection can be a driver of moderate polarisation, which is most pronounced in the attack strategy, while allocentric premise selection has a higher chance of inducing states of lower polarisation, which is most pronounced in the convert strategy.

Table 3.1: Statistics on group divergence in simulation runs as they hit a density of 0.8, based on Leiden clusterings. (†) *Note:* Values for this robustness analysis is based on a sample of 400 simulation runs for each strategy, because not all 1,000 robustness runs reach densities of 0.8. Table reproduced from Kopecky (2022).

Sample	Total		Lowest 10%		Highest 10%	
	Mean	SD	Mean	SD	Mean	SD
Main experiment						
attack	0.41	0.09	0.24	0.05	0.54	0.04
fortify	0.39	0.11	0.17	0.05	0.55	0.03
convert	0.29	0.13	0.08	0.02	0.50	0.04
undercut	0.31	0.14	0.08	0.02	0.54	0.06
any	0.37	0.13	0.12	0.04	0.57	0.04
Agents: 20, Sentence variables: 10 (†)						
attack	0.52	0.11	0.30	0.06	0.70	0.06
fortify	0.49	0.15	0.19	0.06	0.74	0.06
convert	0.30	0.17	0.10	0.00	0.62	0.06
undercut	0.38	0.18	0.11	0.01	0.68	0.05
any	0.40	0.17	0.12	0.02	0.68	0.08
Argument length: 2–4						
attack	0.39	0.07	0.25	0.04	0.50	0.03
fortify	0.34	0.10	0.15	0.04	0.49	0.03
convert	0.27	0.10	0.10	0.02	0.43	0.03
undercut	0.33	0.09	0.16	0.04	0.48	0.03
any	0.36	0.08	0.20	0.04	0.49	0.04

3.3.3.2 Group consensus

When a population of agents is clustered into groups, one can not only ask how much the groups differ, but also how high the agreement is in each individual group. Are they a tightly knit bunch or a more diverse group, in which disagreement may not be uncommon at all? Bramson et al.'s group consensus measures this.

Recall from Chapter 2, Definition 7 (page 45) that group consensus measures the average distance of members of the same group. This measure can capture situations in which the distance in groups is changing over time; contracting groups could be associated with lowering compatibility to outside influences, while rising distance between the group members could indicate that the groups are well acquainted to diversity of opinion and thus more open to outside influence. A rise in group divergence and a simultaneous rise in consensus captures an important part of the intuitive understanding of polarisation.

Figure 3.5 shows how group consensus develops amid the introduction of arguments in the main experiment. It is evident that group consensus correlates with density (Pearson's $r > 0.9$, $p \ll 0.001$ for all strategies), and that the variation between different strategies has a minor effect. Rising group consensus indicates that variance within groups diminish, but it does not automatically indicate that the groups move towards a more extreme stance than initially held by its members, and so the development shown in Figure 3.5 does not quite confirm the law of group polarisation (Myers 1975; Sunstein 2002).

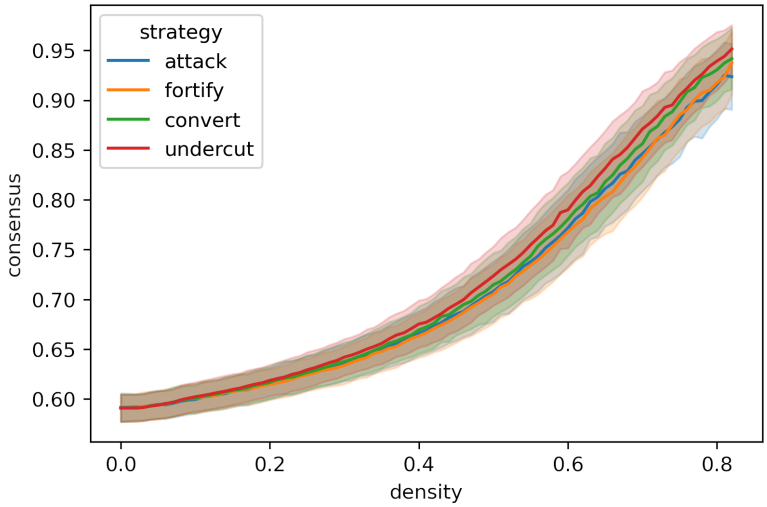


Figure 3.5: Development of group consensus in debates clustered with the Leiden algorithm, depending on argumentation strategy. Shaded areas show standard deviation. Figure from Kopecky (2022).

It seems that the introduction of arguments, virtually irrespective of the employed argumentation strategy, can bring groups closer together. When evaluated together with the results from group divergence, there is a difference in which kind of group is, on average, produced by the argumentation strategies: while convert and undercut arguments lead to groups that are both in internal

agreement and diverge less often from other groups, attack and fortify arguments have a more realistic chance to drive internally agreeing groups further apart, thus generating polarisation.

3.3.4 What about dynamics of opinion diversity?

After observing dynamics of belief polarisation, a natural question to ask is whether argumentation induces dynamics of opinion diversity, too. Unfortunately I was unable to observe any significant effects in this experimental setup. The levels of opinion diversity remain more or less stable throughout most model runs.

3.4 Robustness analysis

Six experiments complement the main experiment, which has a population of 50 agents, 20 atomic sentence variables, and an argument length of 2–3 premises. The first four complementary experiments show that polarisation effects remain at least stable under variation of the initial settings concerning population size, extension of the sentence pool, and length of arguments. Table 3.1 from the previous section shows the mean values for group divergence at a density of 0.8 for the main experiment and two of these robustness analyses.

In a fifth robustness analysis, agents are not initialised with randomly assigned positions, but start off clustered into two groups with perfect bipolarisation. This setting is designed to study the model's behaviour concerning *de*-polarisation rather than polarisation. In the sixth and final analysis, the Leiden clustering is not obtained by taking into account agents' complete positions, but only their stances on four propositions. This analysis is done to accommodate the fact that many real-world debates have a subset of sentences under discussion that are regarded to be the debate's key issues.

3.4.1 20 agents debate 20 atomic sentence variables

The first variant, with 20 agents and the same sentence pool as in the main experiment, shows a slight amplification of the tendencies visible in the main experiment. For example, in group divergence based on Leiden, now 49% of model runs with the attack strategy end with divergences of at least 0.4 (+6%). Conversely, now 25% of runs with the convert strategy result in divergences below 0.2 (+9%). In the other argumentation strategies and in the dispersion measure, existing tendencies in the population of 50 are likewise slightly amplified in the population of 20. Group consensus remains stable.

3.4.2 20 agents debate 10 atomic sentence variables

But it is the second variation, with a population of 20 agents and 10 atomic sentence variables, that shows a considerable rise in polarisation. The middle graph in Figure 3.4 visualises group divergence measurements for debate

stages with a density of around 0.8 in this experiment. The data now accumulates more towards the extremes. In the main experiment, almost a majority of attack and fortify model runs were in the $[0.3, 0.4)$ region. In the run with 10 agents, their largest groups are in the $[0.5, 0.6)$ region. The convert strategy shows a noteworthy shift towards the $[0.1, 0.2)$ region, and the undercut strategy a noteworthy distribution flattening: in the main experiment, its values spike in the $[0.3, 0.4)$ region, but now its data is more smoothly distributed throughout the $[0.1, 0.6)$ interval. In the other three strategies, one can observe the data to spread out in the direction already indicated in the main experiment. As a result, outcomes of high polarisation (≥ 0.8) have a non-zero probability, although still low at maximally 1.5% in the fortify strategy.

This suggests that varying the population size or sentence pool in agent-based debate models with logical constraints does not have a uniform effect towards or against issue polarisation. The effect of these contracting or extending modifications is very much tied to the argumentation strategies, which I count as further evidence regarding their contribution towards polarisation in the model.

3.4.3 10 agents debate 5 atomic sentence variables with arguments of 2 premises

The effect can be even further amplified, as an experiment with 10 agents and five sentence variables shows. Here, the mean group divergence in attack is a staggering 0.67 at densities above 0.65, although the usefulness of this experiment is to be doubted considering that it is quite awkward to imagine a group of 10 agents to debate merely 5 atomic propositions for longer than just a few debate stages. These experiment settings lie on the lower bound of debates that can be feasibly simulated with the present model, as many debates do not reach a density of 0.8, and considering that agents produce arguments with two premises from a total of ten sentences, the argument introduction mechanism must be seen to work under heavy constraints with these settings. On the other end, the upper bound is not characterised by the agents' abilities to devise arguments, but rather by limitations in computational complexity.

3.4.4 50 agents debate 20 atomic sentence variables with arguments of 2–4 premises

It may be more realistic to extend the number of premises that agents may use in devising arguments, up from a length of 2–3 premises that is allowed in the main experiment to 2–4 premises. This raises the pool of premise combinations that are available for argument introduction from 9,120 to 77,520.

The right graph in Figure 3.4 shows the effect of this variation on group divergence at a density of 0.8. There is a contraction of data compared to the main experiment, with a minor amplification in attack and convert model runs: more model runs for attack end in the range for medium polarisation, and more

convert runs end in low divergence (< 0.3). The distributions in fortify and undercut are not amplified, but flattened. Overall, this study seems to confirm the results from the main experiment.

3.4.5 Initial bi-polarisation among 50 agents and 20 sentence variables

In all experiments discussed so far, agents start with randomly assigned positions. This means that the differences between agents is quite homogeneous on average, and so the observed polarisation values are always relatively low at the beginning of a model run. This raises the question how the model behaves when the population starts highly polarised. In this robustness analysis, perfect bi-polarisation is induced by splitting the population of 50 agents in half, and assigning the same position to each agent in each half. All agents in the first group start by assigning True to the first half of sentences, but False to the other half,

$$\{\{p_0, p_1, ..., p_9\} \to \text{True}, \quad \{p_{10}, p_{11}, ..., p_{19}\} \to \text{False}\},$$

and the agents in the second group hold the exact inverse at the start:

$$\{\{p_0, p_1, ..., p_9\} \to \text{False}, \quad \{p_{10}, p_{11}, ..., p_{19}\} \to \text{True}\}.$$

This creates an initial perfect bi-polarisation in terms of group divergence (see Figure 3.6).

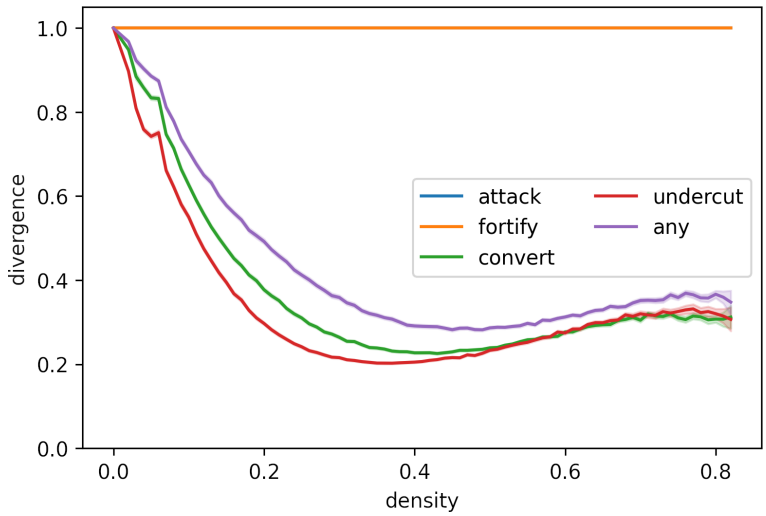


Figure 3.6: Group divergence following Leiden clusterings for an experiment with 50 agents and 20 atomic sentence variables in which the agents start in perfect bi-polarisation. The graphs for the attack and fortify strategies are exactly the same and overlap in this plot. Figure from Kopecky (2022).

There is a striking difference between the argumentation strategies as they respond to initially bi-polarised debates. While the strategies that select premises allocentrically (convert and undercut) show significant effects of de-polarisation, the egocentric strategies (attack and fortify) prove unable to recover from a state of bi-polarisation. Populations that use only these strategies remain at a state of bi-polarisation throughout the debate, while the convert and undercut strategies lead to about the same polarisation values at a density of 0.8 as they did in the main experiment. When allocentrism and egocentrism in premise choice are mixed in the any strategy, the outcome is mixed as well: de-polarisation occurs, but at a lower rate than in the purely allocentric strategies.

3.4.6 Clustering on a subset of propositions

The clusterings in the evaluation above take into account agents' complete positions, which assumes that all sentences under discussion are equally relevant in determining the groups. Yet debates often evolve around a set of key issues. For these, it may be more realistic to cluster agents into groups depending only on their stance towards these key issues. Figure 3.7 shows the results of such a clustering on a subset of the sentence pool consisting in four propositions. The debate stages from the main experiment are used for this analysis, which means that the selecting of "key issues" is much more random compared to how they are distinguished in type 1b models (see Chapter 4). Instead of using agent's complete positions for the clustering, it asks how the population would have been clustered if only these four propositions had been taken into account. The results confirm the main findings, but there is significantly more volatility in the data.

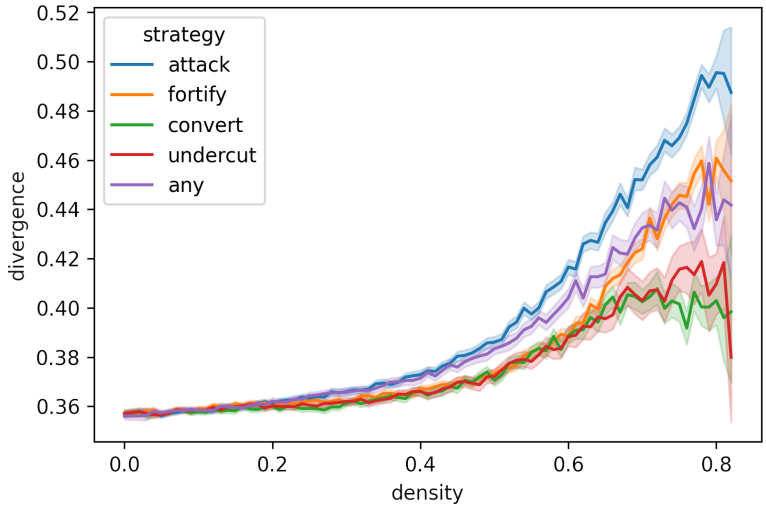


Figure 3.7: Group divergence following a Leiden clustering on the main experiment, but on a subset of four key propositions from the sentence pool. Figure from Kopecky (2022).

3.5 Discussion of results and limitations

3.5.1 Results

In this chapter, I explored simulation experiments on an agent-based, type 1a dialectical structure model. In the experiments, agents update their perspective due to logical constraints from arguments, but not based on social factors such as similarity or trust. A simulation experiment of 5000 model runs revealed that arguments can generally be a driver of issue polarisation, and that argumentation strategies affect it differently. This result was confirmed in a robustness analysis. In dispersion and group divergence, two state-of-the-art measures for issue polarisation, argumentation strategies that behave egocentrically (attack and fortify) in the selection of premises were associated with significantly higher levels of polarisation compared to strategies that select premises allocentrically (convert and undercut). All argumentation strategies increased issue polarisation similarly when observed in a third measure, group consensus. Besides the general influence of arguments on polarisation, the picture that emerged was that the attack and fortify strategies simultaneously lead to groups being more alike internally and more distant compared to other groups, while convert and undercut produced groups that, albeit rising internal consensus, did not move apart from other groups that much.

The argumentation strategies also significantly differed in their ability to recover from bi-polarisation: when agents used the allocentric strategies or a mixed strategy ("any"), they were able to de-polarise debates with initial bi-polarisation. However, the egocentric strategies failed to recover from perfect bi-polarisation and did not show any ability to de-polarise.

The model suggests that polarisation is possible among artificial agents by means of rational processes, although the phenomenon of epistemically rational belief polarisation will be more thoroughly explored in Chapter 4. Argumentation is the sole driver inspected in this model, but it is for future research to inspect polarisation dynamics as argumentation interacts with other factors.

3.5.2 Limitations

The model presented here is intended to understand issue polarisation in a specific kind of artificial agent. The agents always follow the same argumentation and updating strategies, without making any errors in applying them. The results from simulations on this model should not be directly applied in interpretation of human behaviour and/or states of social polarisation. Rather, this model elucidates properties of argumentative features irrespective of other variables, of which there are quite a few.

This model fills one, but not all gaps mentioned in the literature. Polberg & Hunter (2018) stress the importance of modelling (a) bipolar argumentation, allowing for both support and defeat relations in agent-based models, but also of modelling (b) probabilistic belief systems. The model presented here fulfils

their requirement (a) but falls short of fulfilling (b), mainly due to computational restrictions. An extension of the model to probabilistic belief systems is left for future research.

As mentioned above, the simulation results fall short of producing high and very high polarisation values. This is in contrast to some social influence models (Banisch & Olbrich 2021; Mäs & Flache 2013), which often end in states of perfect bi-polarisation. Yet this inability to produce perfect bi-polarisation should be seen as a virtue rather than a vice. If argumentation alone were to explain high and very high degrees of issue polarisation among artificial agents, there would be no room to accommodate other factors in extended models. The factors not considered in this model include homophily, limited agent memory, and bias in selection of communication partners relative to argumentation strategy. Extensions of this model could consider if there should not be some bias in selecting a target position given some of the argumentation strategies: for example, what changes if agents only attack out-group targets?

3.6 Supplementary materials

A repository at https://zenodo.org/records/5523209 contains the complete simulation runs including data on positions and arguments as pickled Python objects, raw measurement values for each polarisation measure, as well as the raw ARI values, as zipped DataFrames, Jupyter notebook for the simulation experiments and data analysis, as well as the source code and documentation of `taupy` in the version used in this chapter.

Chapter 4

Argumentation-induced rational issue polarisation

4.1 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 towards one's own views and a disconfirmation bias towards 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 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.

The limitation to agent memory in Singer et al.'s model are quite severe and 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 Section 4.2.

Dynamically evolving dialectical structure models help remedy this situation. In type 1b dialectical structure models, 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 (Section 4.3). This possibility raises new questions about what, if anything, is wrong with issue polarisation, and which interventions its occurrence requires (Section 4.4).

The results go beyond that. Simulations on the type 1b dialectical structure 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 Section 4.4, these results underpin the relevance of argumentation to matters in social epistemology.

4.2 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. They build their case from simulation experiments on a computational model that captures multi-agent deliberative processes, such as in *Twelve Angry Men*, a 1955 play by Reginald Rose. This play 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. It is all the more surprising that belief polarisation should arise in these circumstances.

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 (see Figure 2.20).

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 $\emptyset$. 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*, 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 & 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).

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

4.3 Simulation procedure and results

4.3.1 A brief description of the model in use

This chapter relies on a type 1b dialectical structure model, an extended version of the model used in Chapter 3. There are twelve agents in each model run to capture scenarios of deliberation like in *Twelve Angry Men* and Singer et al. (2019). As in type 1a models, a debate evolves through agents' introduction of arguments. But unlike the random debates that type 1a models create, the argument introduction process in type 1b models ensures that argument maps grow in a tree-like fashion, as illustrated in Figure 2.16. Some elements of the sentence pool in type 1b models are designated to be the key propositions of the debate. In *Twelve Angry Men*, one of the key propositions would be that the defendant was guilty. Key propositions are conclusions of arguments at the root of argument maps. The sentence pool in type 1b models is also expandable. In some of the simulations reported below, the agents are initially aware of 15 atomic propositions. Five additional propositions are introduced in the course of model runs. Agents hold not just binary views in type 1b models, but can also withhold judgement for any of the propositions under discussion.

The argument introduction and belief updating processes are described in Chapter 2. In each model run, all agents share the same introduction strategy and only introduce arguments according to this strategy, even though the any strategy allows for variation in behaviour. Belief updating following proposition pool extension and argument introduction ensures that this model can be recognised as showing epistemically rational behaviour. I justified this claim in Section 2.4.1.4.

An overview of simulation procedure for type 1b models is shown in Figure 2.15. These simulation experiments are performed with the goal of supporting the case for epistemically rational belief polarisation.

4.3.2 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. This is the conjunction of features 1 and 2 in Esteban & Ray (1994, p. 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.

The partitioning into groups, or simply "clustering", required to calculate these values follows the description in Section 2.3.2. It is obtained through two state-of-the-art community structuring algorithms for social networks, Leiden (Traag, Waltman & van Eck 2019) and affinity propagation (Frey & Dueck 2007). Using multiple algorithms is one strategy to verify that the obtained clusterings are reliable.

4.3.3 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 Section 4.3.4, antecedently polarised in Section 4.3.5 and initialised with 80% agreement in Section 4.3.6. Sections 4.3.5 and 4.3.6 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*. Recall from Section 2.3.5 that 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, p. 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 4.3.7 varies this termination condition by looking at the dynamics beyond $D = 0.8$.

4.3.4 Results from randomly allocated belief systems

Figures 4.1 and 4.2 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 before the modelled debate commences, but several robustness analyses in Section 4.3.5 and 4.3.6 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 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, agent's 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.

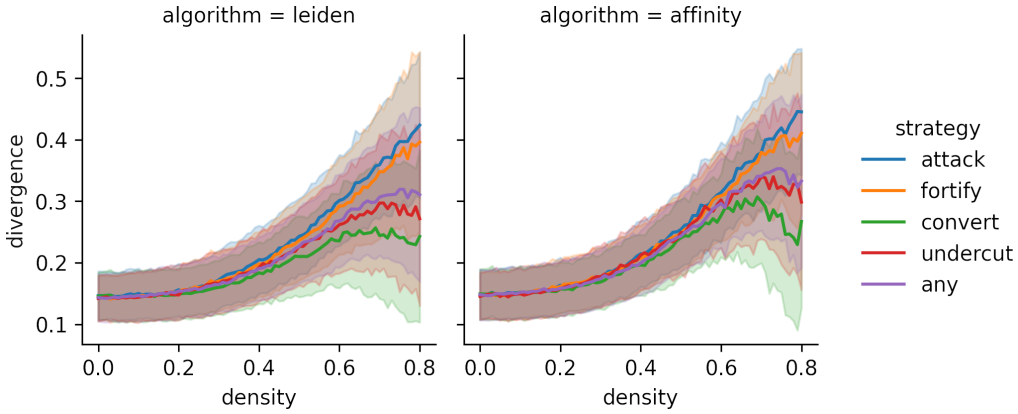


Figure 4.1: Group divergence dynamics from two clustering algorithms under starting condition of low polarisation (completely random position assignment). The mean of 1,000 runs per strategy is shown in the line plot, and the data's variation of ±1 standard deviation is plotted in the adjacent shaded area (Figure from Kopecky 2024).

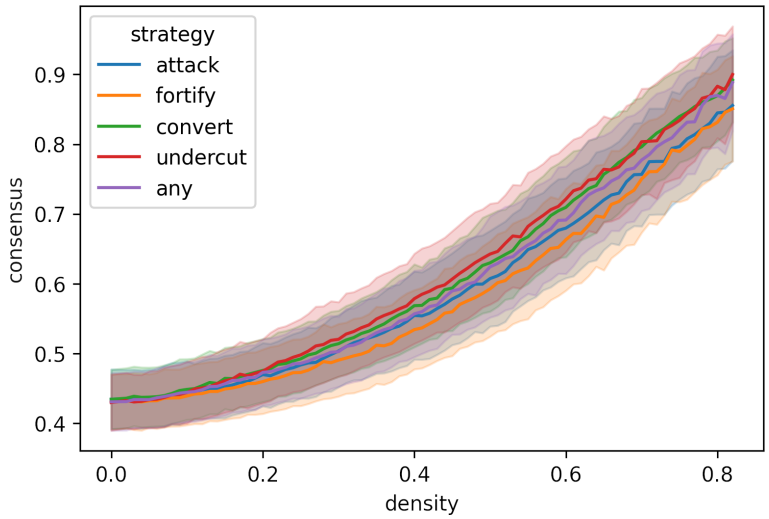


Figure 4.2: Group consensus dynamics from Leiden clusterings under starting condition of low polarisation (completely random position assignment). The mean from 1,000 model runs per strategy is plotted as a line, and variation of ±1 standard deviation is indicated by the adjacent area (Figure from Kopecky 2024).

The difference between the argumentation strategies is particularly visible at a density of $D = 0.8$. The group divergence attained by attack and convert model runs at this point is shown in Figure 4.3. It indicates that agents are depolarised in the convert model runs but significantly more polarised when they follow the attack strategy. These two strategies seem to induce very different deliberative settings at higher degrees of inferential densities.

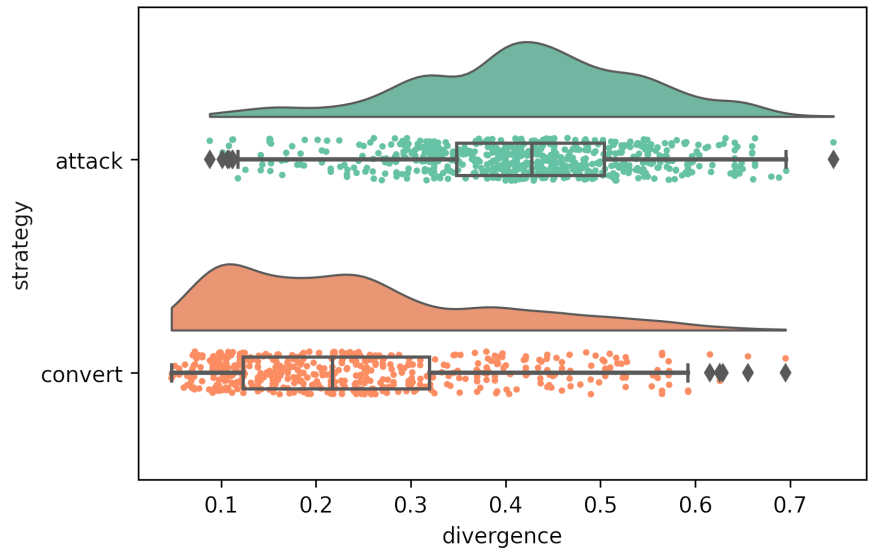


Figure 4.3: Distribution of group divergence results following Leiden clusterings at a density of $D = 0.8$ for two strategies, attack and convert (Figure not previously published).

The different results for different argumentation strategies are interesting because they cut 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 (Section 4.4).

4.3.5 Results from antecedently opposed beliefs

Agents starting a debate with 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 Figure 4.4 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 Table 4.1), which eventually expanded to 20. Four agents were assigned to each group.

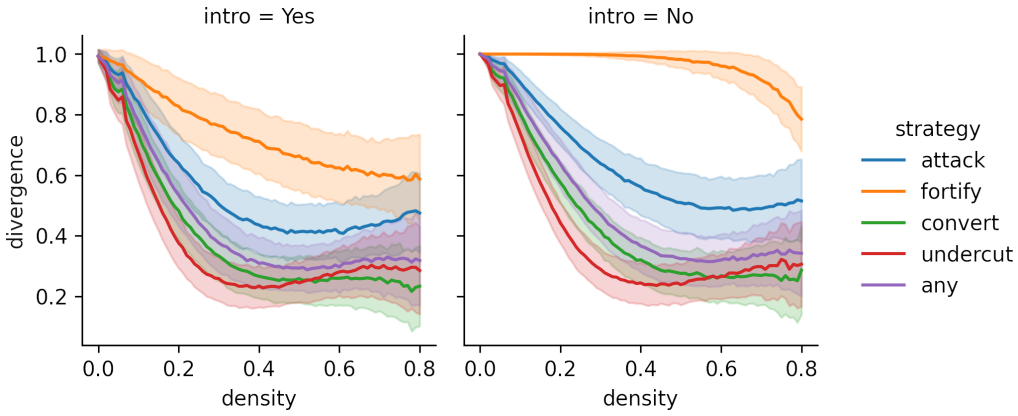


Figure 4.4: Group divergence dynamics from 1,000 runs per strategy following Leiden clusterings with groups antecedently configured to have maximum polarisation. The left, but not the right plot, shows experiments with proposition pool expansion. Means and standard deviation are indicated as before (Figure from Kopecky 2024).

Table 4.1: Initial beliefs held by antecedently polarised agents

Group 1	Group 2	Group 3
$p_0, ..., p_4$ → True	$p_0, ..., p_4$ → False	$p_0, ..., p_4$ → None
$p_5, ..., p_9$ → False	$p_5, ..., p_9$ → None	$p_5, ..., p_9$ → True
$p_{10}, ..., p_{14}$ → None	$p_{10}, ..., p_{14}$ → True	$p_{10}, ..., p_{14}$ → False

In the previous experiment, the sentence pool expanded by a third of its original size. This feature is absent in the second scenario (Figure 4.4, right), where 21 propositions in the sentence pool were known to the agents initially (Table 4.2) 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.

Table 4.2: Initially polarised beliefs that span the entire sentence pool

Group 1	Group 2	Group 3
$p_0, ..., p_6$ → True	$p_0, ..., p_6$ → False	$p_0, ..., p_6$ → None
$p_7, ..., p_{13}$ → False	$p_7, ..., p_{13}$ → None	$p_7, ..., p_{13}$ → True
$p_{14}, ..., p_{20}$ → None	$p_{14}, ..., p_{20}$ → True	$p_{14}, ..., p_{20}$ → False

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.

4.3.6 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 4.5 (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 4.5 (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.

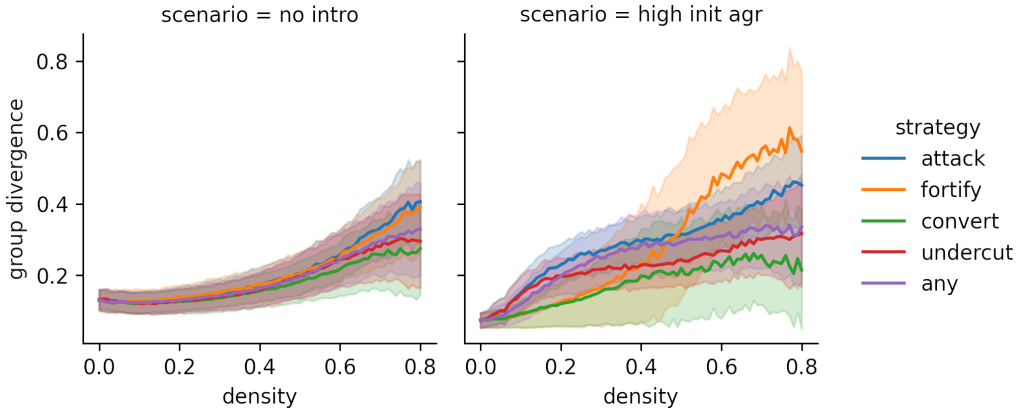


Figure 4.5: Mean group divergence determined through Leiden clusterings for 1,000 simulation runs per strategy. These are variations of the base experiment in which 20 sentences are known initially and no new sentences introduced (left) and with initially highly agreeing (80%) populations and a fully known sentence pool (right). The shaded area displays values ±1 standard deviations away from the mean, which is indicated by lines (Figure from Kopecky 2024).

In the initially polarised populations in Section 4.3.5, the introduction of additional sentences influenced the results and particularly affected the fortify strategy. In initially random beliefs, however, no noticeable deviance from the original results can be observed (Figure 4.5, left, note the different y axis scale compared to Figure 4.1).

Polarisation dynamics are observable even if agents initially agree on most sentences (Figure 4.5, 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 (Figure 4.4), it also appears able to break up agreement among highly agreeing agents. Beyond the polarising fortify strategy, Figure 4.5 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.

4.3.7 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.

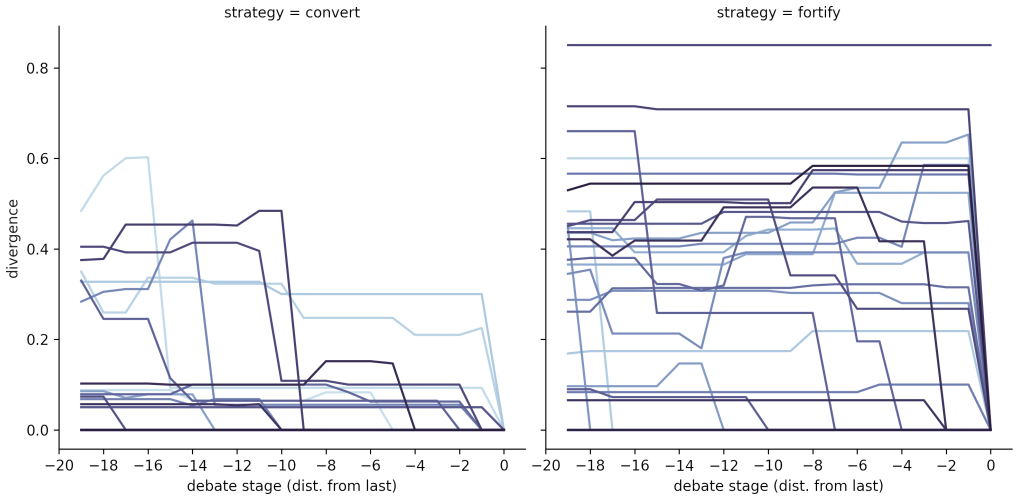


Figure 4.6: Group divergence following Leiden clusterings in 40 randomly sampled convert and fortify runs that continue until the maximum inferential density. The single fortify run that does not collapse is an example where termination occurs before the maximum density is reached, presumably due to unsuccessful argument introduction (Figure from Kopecky 2024).

Argumentation strategies differ in how agents approach this ideal point. Figure 4.6 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.

4.4 Discussion: 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 – for example, Popper (1976, p. 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). 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 towards 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 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 towards 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?

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 & 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.

4.5 Supplementary materials

A repository at https://zenodo.org/record/6448599 contains Jupyter notebooks to run the experiments described in this chapter and analyse the data.

Chapter 5

Inconsistent belief aggregation in diverse and polarised groups

Note

The narrative voice changes from singular "I" to plural "we" in this chapter. While this breaks continuity with the rest of this thesis, it highlights that this work was gathered from collaborative research (Kopecky & Betz 2025).

5.1 Introduction

When citizens and their governments rely on expert groups for policy advice, should they favour a diversified group composition? Studies from models of epistemic landscapes and network epistemology suggest an affirmative answer. Groups with diverse viewpoints cover a high proportion of the approaches to a problem (Grim et al. 2019; Hong & Page 2004), 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, type 2 dialectical structure model (Section 2.4.2), 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, p. 203) to obtain a representative group opinion quickly when consensus can not 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 & Pettit 2002) – a prime instance of the independence phenomena introduced in Chapter 1.

In simulations on our model, we observe inconsistent majoritarian aggregation predominantly in highly diverse, 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 can not 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 chapter begins with a description of foundational concepts in Section 5.2. There we describe majoritarian belief aggregation and the inconsistent group beliefs that it is prone to. We then describe our agent-based model and present simulation results in Section 5.3. We reflect on these results and discuss possible consequences for expert advice in Section 5.4.

5.2 The independence phenomenon arising in majoritarian belief aggregation

5.2.1 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, 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 assume that uncertainty surrounds the issue and 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 & Pettit 2002). Table 5.1 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.

Table 5.1: 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

Inconsistent belief aggregation is a problem for groups that issue recommendations to the public and policy makers. In the following Section 5.2.2, 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.

5.2.2 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 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 & Pettit 2002, p. 93; Pettit 2001, p. 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 in case 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, pp. 279–280). 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 5.1. 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, p. 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 later (Section 5.4).

The goal for this chapter 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. This question can be pursued with the diversity and polarisation measures from Section 2.3 as well as the algorithms that synthesise agent samples in epistemic decision problems with varying number of arguments and specific levels of diversity and polarisation (as described in Section 2.4.2).

5.3 Simulation procedure and results

5.3.1 Epistemic group decision problems in type 2 dialectical structure models

How likely are diverse and polarised groups to aggregate inconsistent group opinions through sentence-wise majority voting? We apply an agent-based, dialectical structure model with synthesised argument maps to pursue this question. To recall from Section 2.4.2, these models consist of two sub-processes. The first sub-process generates a synthetic argument map as the basis of the group's decision problem (Section 2.4.2.1). This dialectical structure is hierarchical with arguments on the lowest level leading to conclusions that are the key propositions of the decision problem. Arguments can defeat and support each other, and not every belief system is admissible in light of the argument map. The second sub-process (Section 2.4.2.2) samples agents to generate a group with arbitrary degrees of opinion diversity and belief polarisation, as understood by the measures from Section 2.3.

Diversity is understood in terms of the Gini–Simpson diversity index introduced in Section 2.3.4. We found this choice to be justified given both the group size of less than 100 individuals and the particular aggregation procedure that we study. In small groups, the Gini–Simpson index is a more reliable indicator of diversity compared to other relative frequency approaches such as the Shannon index (Tuomisto 2010, p. 854). Regarding the sentence-wise majority voting aggregation procedure that we study, we realised that a diversity measurement in terms of absolute frequency, or richness, could easily yield misleading results. Consider that in a group of 51 agents, the addition of a single agent can double the group's richness even though the addition of a single vote is not likely to influence the outcome much. Richness would suggest a fundamental change to group composition, but this would contrast with the comparatively small influence we expected individuals to have in majority voting. For these two reasons, heterogeneity understood as the Gini–Simpson index seems the most reliable indicator of diversity for the present purpose.

The argument maps that form the basis of the decision problems that we study 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. We studied decision problems with varying inferential density, and arguments are added iteratively to the maps until a specified value of inferential density is reached (see Section 2.3.5).

The example from Table 5.1 in Section 5.2.1 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 5.2.1, such as the one in Figure 5.1.

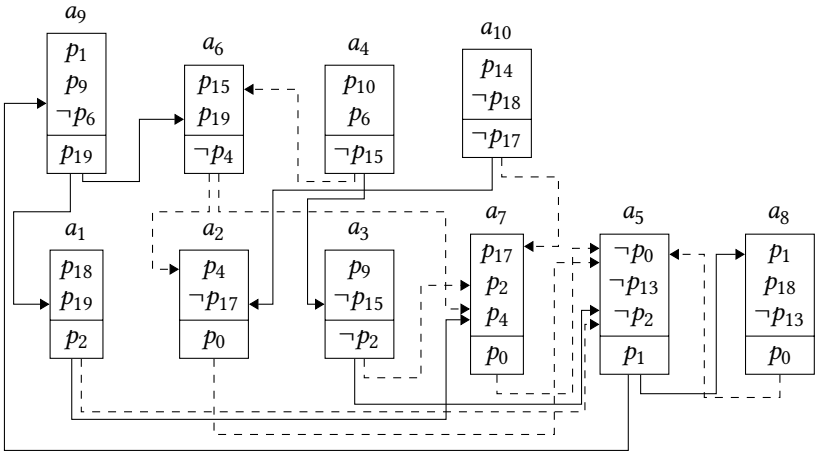


Figure 5.1: Illustration of a synthetically generated argument map with key statements p_0, p_1, and p_2. Support relations are expressed by solid arrows, defeats by dashed ones.

From agents with beliefs that are characterised in this way, our model samples groups with a specified degree of diversity or polarisation, depending on the model variant in use. 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 (see Section 5.5), but we include a brief summary here. For the diversity variant of the model, we first apply the affinity propagation clustering algorithm (Frey & 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. That is, we ask whether the group opinion aggregated from individuals that respect the validity of all presented arguments in turn respects all presented ar-

guments. 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 5.3.2, we discuss the model parameters and summarise the main results, and we present the results in more detail in Section 5.3.3.

5.3.2 Model parameters and main results

In our model, 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 none inconsistent aggregation. In our explanation for this initially counter-intuitive pattern (Section 5.3.4), 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 are 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 case of a tie. From Betz's formula (2013, p. 44), we determine the number of validity-respecting beliefs at each inferential density D by solving for x in the equation $D = (20 - \log_2(x))/20$. At a density of $D = 0.4$, 20 propositions allow for 4,096 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.

5.3.3 Quantitative explorations of many model runs

In this section, we present results from thousands of iterations of both the diversity and polarisation variant of our model.

Figure 5.2 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 5.3 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 5.3.4.

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 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 5.2 and 5.3 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 5.2. It illustrates that, in our model, the prevalence of inconsistent majoritarian aggregation continuously rises as groups diversify. Groups with both very high and very low polarisation are likely to achieve consistent

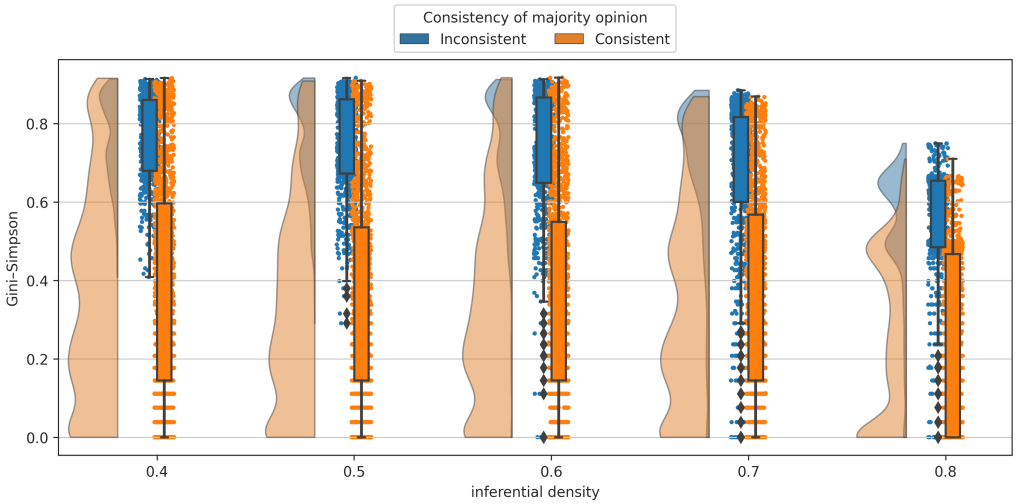


Figure 5.2: Majority opinion consistency in 10,798 samples of 51 agents with varying diversity, expressed as the Gini–Simpson index, and varying informational influence, expressed as inferential density. Scatter plots show all observations, while the box plots indicate the data points within the 25th to 75th percentile. 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 from Kopecky & Betz (2025).

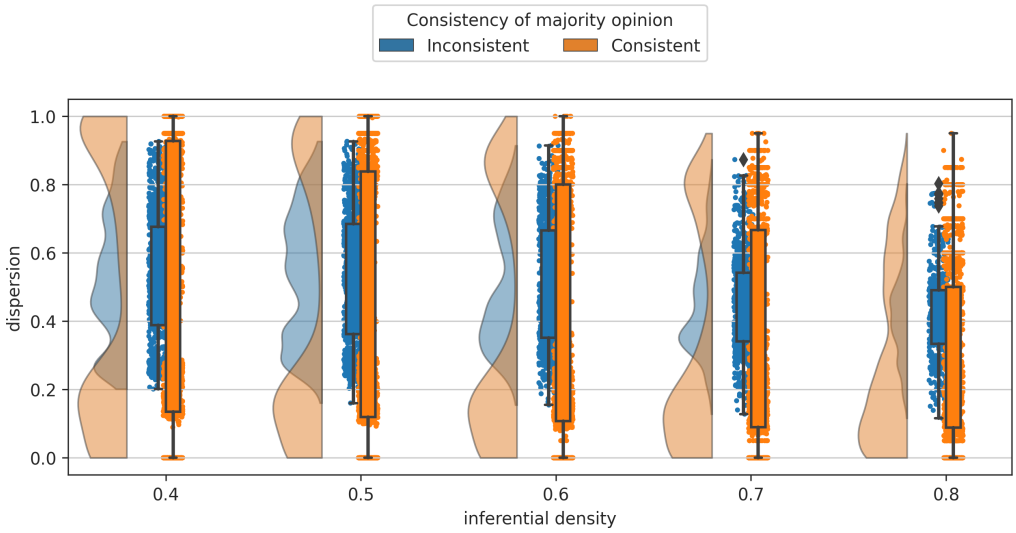


Figure 5.3: Majority opinion consistency in 10,722 samples of 51 agents with varying polarisation, measured as dispersion, and inferential density. See Figure 5.2 for further description. Figure from Kopecky & Betz (2025).

majoritarian aggregation, but moderately polarised groups achieve it in less than a third of all cases.

Table 5.2: Ratio of consistent group beliefs, across all points of inferential density, depending on diversity (left) and polarisation (right). 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

These two effects can be further quantified using a binary logistic regression analysis. With consistency as 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 5.2 and 5.3, Cohen's f^2 indicates a strong effect for diversity (0.40) and an even stronger one for polarisation (0.54).

5.3.4 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, 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 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, p. 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 bi-polarised populations, by definition, form two clusters. This clustering dissipates as groups diversify, leading to more clusters that have less members (as displayed in Figure 5.4).

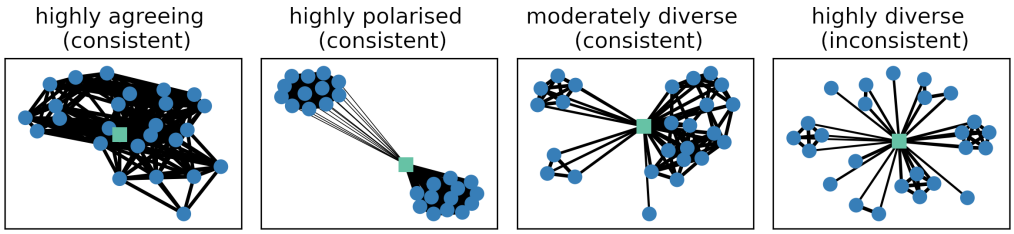


Figure 5.4: Four samples with 25 agents each. The majority opinions are printed as green squares and the agents as blue circles. Relative node position is a rough indicator of distance. All edges between agents and the majority opinion are plotted and weighted by distance, but edges between agents are 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 1 cluster and the bi-polarised sample has two clusters. As opinion diversity rises in the group, more and more clusters become discernible until none can be made out. Figure from Kopecky & Betz (2025).

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 5.3 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

Table 5.3: Illustration of a sub-group $a_1, ..., a_j$, $j > m/2$, determining the majority's position on propositions $p_1, ..., p_i$ 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	...	p_i	p_{i+1}	...	p_n
a_1	+	+	...	+	?	...	?
a_2	+	+	...	+	?	...	?
⋮	⋮	⋮	...	⋮	⋮	...	⋮
a_j	+	+	...	+	?	...	?
a_{j+1}	?	?	...	?	?	...	?
⋮	⋮	⋮	...	⋮	⋮	...	⋮
a_m	?	?	...	?	?	...	?
Majority	+	+	...	+	?	...	?

well, by mere statistical likelihood. This likelihood of achieving consistency by chance drops as fewer opinions remain consistent at higher inferential density.

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.

5.4 Discussion: 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 5.1.

As inconsistencies might escape an external party, such as in the *expert poll* scenario outlined above, we can not recommend to pursue 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 & 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 proportion 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 & 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 can not 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.

5.5 Supplementary materials

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

Chapter 6

Conclusion

Humans do not always agree about how to respond to a body of evidence. When they engage in deliberation, problem-solving in groups, and the pursuit of knowledge, a linear progress and convergence may fail to materialise even if all participants share the same goals, a productive mindset, and meet standards of epistemic rationality. Phenomena associated with this mismatch between individual rationality and collective optima can be called *independence phenomena*. In this dissertation I have studied independence phenomena in epistemically rational, artificial agents. I studied cases in which individually rational behaviour could give rise to outcomes that are collectively suboptimal, such as states of belief polarisation and inconsistent majoritarian belief aggregation. These phenomena could be described within the theory of dialectical structures, an argumentation-based framework. The study of argumentation in this framework turned out to be remarkably fruitful for our understanding of social-epistemic dynamics.

The results in this dissertation describe social-epistemic phenomena, but do they also provide normative guidance for individual epistemic behaviour? If you knew that reasoning allocentrically would depolarise your community, should that compel you to reason in this fashion? This question is difficult to answer, because it is unclear whether individuals can be required to contribute to depolarisation of belief in their community. Maybe some transient polarisation of belief is necessary to overcome established dogma? Likewise, it is not obvious that avoiding expert panels with diverse opinions to ensure consistent recommendations is always the way forward. It might not even be a practicable option if high diversity is a true reflection of expert opinion on the current matter. In these questions and phenomena, the focus of this dissertation is mainly on understanding epistemic circumstances we may find ourselves in.

Simulation experiments on dialectical structure models give further support for the thesis that agents can polarise in their beliefs through interaction with others even if they adhere to epistemic rationality standards. I found that epistemic rationality does not necessarily prevent polarisation of belief – not trough empirical observation but as a matter of principle. Memory limits or trust networks need not be assumed to observe this phenomenon. The simulations further show that there is a substantial difference in the impact of egocentric versus allocentric strategies in multi-agent argumentation and reasoning. Agents that continuously gather evidence in favour of their own theory deserve praise for improving highly consistent beliefs and theories. But

when they do not engage with the beliefs of others, they are much more likely to polarise their beliefs. This observation extends earlier results about the effect of argumentation on consensus formation and the likelihood that groups attain true beliefs through argumentation (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 belief 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.

This first group of results was obtained from dialectical structure models in which argument maps evolve through iterative argument introduction. Dialectical structure models can also be used for a second type of agent-based model. In this second type, argument maps are synthesised and groups of agents that respond to these maps are sampled. I used this second type to ask whether groups with highly diverse beliefs are always better in responding to epistemic group decision-problems, and polarised groups always worse? The answer was: 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.

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 ill 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 incompatible approaches. 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 can not 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 case of inconsistent or inconclusive recommendations, such as by over-arching agreement on cultural or moral ideals?

The results from dialectical structure models 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. Dialectical structure models are not simple ones and yet they yield results that perfectly complement the knowledge obtained from models in other approaches. We should be looking forward indeed to the new questions that computational philosophy will be able to tackle in more ambitious models.

Appendix A

A description of the base model in the ODD protocol

This appendix describes the type 1a dialectical structure model according to the ODD protocol (Grimm et al. 2006). It was previously published as Section 3 of Kopecky (2022). The ODD protocol is a standard protocol for documenting and describing agent-based models. The protocol is designed to describe agent-based models independent of the conventions of any one academic discipline. It can thereby help modellers of different academic backgrounds in understanding and implementing models designed by others.

A.1 Overview

A.1.1 Purpose

This model is designed to study the effects of argument introductions and argumentation strategies on issue polarisation in debates among artificial agents.

A.1.2 State variables and scales

Debates in this model are simulated as the logical conjunction of arguments. Arguments are logical implication relations between a set of premises and a conclusion. Both premises and conclusions are drawn from a sentence pool, which consists of n atomic sentence variables and their negations, meaning it has $2n$ elements. Every argument introduced to the debate must meet these criteria:

Satisfiability: A debate must remain satisfiable at all times. That is, the conjunction of arguments must be satisfiable.

Premise uniqueness: Every argument must have a unique set of premises, i.e. any sets of premises can be used in at most one argument of the debate. This restriction does not hold for conclusions, i.e. there can be multiple arguments with the same conclusion.

Prohibition of conflicts and redundancy: If a sentence is used as a premise, neither it nor its negation is used as the conclusion or another premise of the same argument.

A debate is represented by a Boolean expression of the conjunctive form in (A.1).

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

The arguments and the debate as a whole use sentence variables, and the model is thus abstracting from the actual propositional content of premises and conclusions. This is sufficient for the purpose of investigating the general role of argumentation in polarisation dynamics, but building debates on sentence variables can not elucidate the role of particular propositional contents in premises (such as the difference between normative and descriptive claims) or of actual argumentation schemes. Natural language argumentation technologies seem necessary for this task, and it will be interesting to see how emerging approaches to natural language processing of argumentation (Betz 2022; Hunter et al. 2019) can be employed in future research.

Agents in this model are simulated as having a belief system represented by positions in terms of TDS, and I will use both terms interchangeably in this paper. Positions are mappings from the atomic sentence variables to truth values True and False. An agent's belief system is fully specified by these truth-value attributions. In this model, agents assign a truth value to every sentence in the sentence pool (they never suspend judgement), but are confined to satisfying interpretations of the debate, which means that every agent must hold a position that is an interpretation of the Boolean formula that describes the debate. This minimal picture of rationality implies that agents assign identical truth values to equivalent sentences but different truth values to contradictory sentences, and follow their inferential obligations: if an agent assigns True to all premises in an argument, it also assigns True to the conclusion.

For a simulation with n sentence variables in the sentence pool, an agent's position can be represented as in (A.2):

$$\begin{aligned} p_1 &\rightarrow \text{True} \\ p_2 &\rightarrow \text{False} \\ &\vdots \\ p_n &\rightarrow \quad ... \end{aligned} \tag{A.2}$$

Agents that assign True to an atomic sentence variable are said to "accept" it, otherwise they "reject" it. Besides a so-defined belief system, agents are associated with one of the five argumentation strategies (described as part of the argument introduction sub-process below). In every model run, all agents share the same argumentation strategy. Distances between positions of any two agents are measured by means of the Hamming distance, which is interpreted to be the number of sentences that are mapped to a different truth value.

A debate stage is described by the debate at that time (the current state of the conjunction of the arguments) together with the agents' current positions. There are a number of high-level properties that can be obtained from these

Table A.1: Description of the model parameters and values adopted in the main experiment

Parameter	Value
No. agents	50
No. atomic sentence var.	20
No. premises per argument	2 or 3
Argumentation strategy	Either of five
Updating strategy	Closest coherent
Initial positions	Random
Max. density	0.8

lower-level properties. The first object of interest here is the argument graph. An argument graph is a two-coloured directed graph that takes the arguments in the debate as nodes and defeat and support relations as edges. A pair of arguments (a, b) satisfies the *support* relation if the conclusion of a is equivalent to one of the premises in b, and the pair fulfils the *defeat* relation if the conclusion of a is equivalent to *the negation of* one of the premises in b. This means that the relations between arguments are automatically obtained from the arguments. Argument graphs are not necessarily complete, and are non-circular more often than circular. The argument graph of a debate stage i is referred to as τ_i.

A second group of higher-order properties of a given debate stage concern its *space of complete and coherent positions*, represented as Γ_τ (SCCP, Betz 2013, pp. 39–41). In logical terms, the SCCP is the set of all satisfying interpretations of the Boolean formula that represents the debate at a given stage (see Figure A.1 for an example). It should be noted that the SCCP is very different to, and usually contains much more elements than, the collection of actually held positions by the simulated agents. Actually held positions have to be in the SCCP, but multiple agents can hold the same position from the SCCP, and the actually maintained positions in the simulation can be quite spread out in the SCCP. In terms of the model, the SCCP is the set of positions that the actual agents are allowed to move to should their positions be rendered incoherent by the introduction of an argument. A position is incoherent relative to a debate stage if its assertions are jointly unsatisfiable with the arguments at that debate stage. Other than that, the model allows agents to move freely in the SCCP. In particular, it does not prescribe them to favour positions with maximal quantitative argument support, or to move toward positions held by other agents. This seems realistic considering that it can be rational to adopt a position even if there is just a single argument in its favour, namely when the single argument is especially convincing. Argument evaluation and a measure of argumentative strength are not part of this model, however.

The size of the SCCP, $|\Gamma_\tau|$, is used in calculating a debate stage's *density*, a fundamental measure of progress in debate simulations (Betz 2013, pp. 44–49). Roughly speaking, density encodes how many positions have been rendered

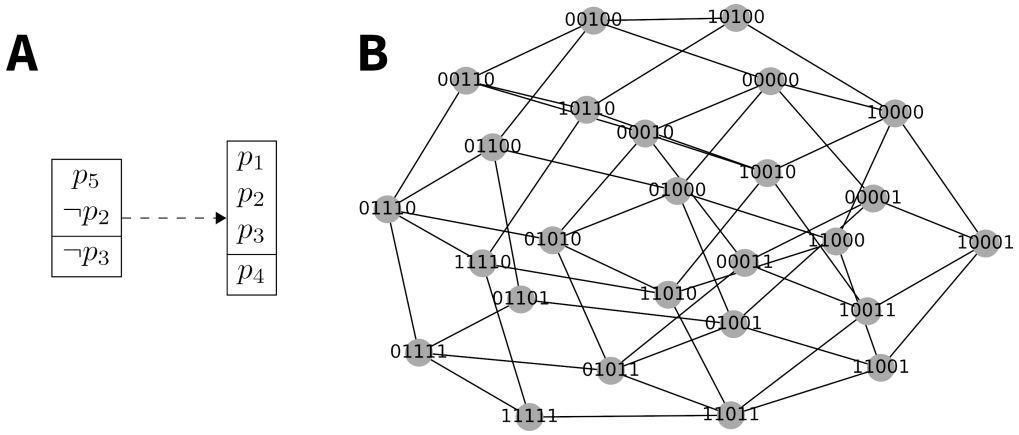


Figure A.1: (**A**) Argument graph for the debate stage $((p_1 \wedge p_2 \wedge p_3) \Rightarrow p_4) \wedge ((p_5 \wedge \neg p_2) \Rightarrow \neg p_3)$, and (**B**) the resulting space of coherent and complete positions (SCCP). In the argument display, premises and conclusions are separated by a horizontal bar, and the defeat relation is expressed with a dashed arrow. In the SCCP, nodes have a label that shows the bit string representation of the position they resemble. In this string, sentences are ordered alphabetically (p_1 will show in the first bit, p_5 in the last), and they are either 0 if the position assigns False to this proposition, or 1 if the position accepts the proposition. Positions are connected by an edge if they differ in exactly one truth-value attribution (i.e., if their Hamming distance equals 1).

incoherent so far in the debate, and how much freedom the agents have in choosing the next position to move to if they have to. Importantly, not every argument introduction raises the debate's density in the same way, and debates can take up to twice as much argument introductions to reach the same density. A debate stage's density is always in the $[0, 1]$ interval, and defined as $(n - \log_2(|\Gamma_\tau|))/n$, where n is the number of atomic sentence variables (Betz 2013, p. 44). Figure 12 in Kopecky (2022) shows how density evolves over simulation time depending on the different argumentation strategies.

The most important parameters of the model and initial settings for the main experiment are described in Table A.1. Almost all of them influence the simulation's computation time. The number of premises per argument is initially chosen as the range 2–3, but a robustness analysis was run on 2–4 sentences per argument. This parameter can be either a number or a range. If it is a range, the number of premises in any introduced argument is chosen randomly. The number of atomic sentence variables and the termination density are taken as best practices from Betz (2013). The number of agents are chosen somewhat arbitrarily, but robustness analyses are provided for smaller populations and sentence pools (which actually confirm and exceed the results). Varying the size of the sentence pool does exponentially influence the run time of experiments. Given the current software implementation and the available computational resources for conducting the experiments reported below, a size of 20 atomic sentence variables, resulting in a sentence pool of 40 sentences, proved workable.

A.1.3 Processes and scheduling

The simulation proceeds by two kinds of events: argument introduction and a following position updating, and simulation time is understood as numbers of introduced arguments. At every step, both events are called until limiting conditions are met. The model terminates either when a density greater or equal than the maximum density parameter is reached, or if an argument introduction fails due to lack of premises or conclusions meeting the requirements imposed by the strategy.

For each argument introduction, two agents are randomly drawn from the population. The first agent is understood to be the *source*, the second one is called *target*. The source then acts according to its associated argumentation strategy. A single model run assigns the same strategy to all agents, which can either be one of the four basic ones (Betz 2013, pp. 93–94):

Attack: From the sentence pool, the source picks premises that it accepts and a conclusion that the target rejects to build a valid argument. The source also ensures that the conclusion does not contradict its position.

Fortify: From the sentence pool, the source position selects both premises and a conclusion that it accepts to construct a valid argument.

Convert: From the sentence pool, the source position selects premises that the target accepts and a conclusion that the source accepts to build a valid argument.

Undercut: From the sentence pool, the source position picks premises that the target accepts and a conclusion that the target does not accept to construct a valid argument. The source also ensures that the conclusion does not contradict its position.

Agents can also have a fifth strategy, which picks a strategy at random for each argument introduction:

Any: The source randomly chooses one of the four basic argumentation strategies to introduce a valid argument.

When this argumentation process is completed, all agents in the population check whether their positions are rendered incoherent by the new argument. From the logical point of view, an interpretation can satisfy one Boolean formula, but not the updated conjunction of the Boolean formula and an added formula (i.e., the previous debate extended by a newly introduced argument). When this happens, a position is rendered incoherent. In the model, all the agents with incoherent positions following argument introduction immediately update their position. Figure 11 in Kopecky (2022) shows how many agents, on average in the model runs of the main experiment, update their position following argument introduction. As can be seen, the undercut strategy has the agents update considerably more often than the other strategies. As the SCCP shrinks and density rises accordingly, there is comparatively high pressure on the agents to update their positions. After all, a shrinking SCCP implies that a decreasing number of positions are acceptable to the agents.

The position update strategy that all agents share points them to their respective closest coherent position from the SCCP. "Closest" is understood to mean lowest Hamming distance. When there are several coherent positions with minimal Hamming distance, one of them is chosen randomly. Distances to the positions held by other simulated agents do not influence the updating process.

A.2 Design concepts

After this technical review of key objects and processes, let me provide a more conceptual reflection of the model's properties.

Emergence: The starting positions of agents are randomly assigned at the start of each model run. However, agents autonomously select their subsequent updated positions based on coherence criteria. Levels of polarisation in this sense emerge from the model.

This is also true for the relation between propositions: whether an agent can simultaneously accept two items from the sentence pool depends on the autonomously introduced arguments.

Adaptation: Agents update their position if a newly introduced argument renders their position incoherent. They do so by moving to the closest neighbour among the remaining coherent positions in the SCCP, or selecting a random next neighbour if more than one have minimum Hamming distance to their previous position. In this way, only the logical relations matter for an agent's adaptation.

Although more than one agent can hold a coherent position, agents select the closest position among all coherent positions, not just those that are currently held by other agents in the simulation. If updating is required, they choose one of the coherently adoptable positions without regard for what others believe, even their closest neighbours.

Fitness: Agents have only two goals. The first is upholding a coherent position, which they maintain by the adaptation process just described. Their second goal is to introduce arguments according to their assigned argument strategy if the model determines that it is their turn. Agents always fulfil both goals: since there is no distance limit in selecting a new position, updating always succeeds for every agent. Also, the simulation stops when one agent is unable to introduce an argument according to its argumentation strategy. Polarisation is thus not introduced due to agents' inability to accomplish their goals.

Sensing: Agents that are selected for argument introduction know their own position and the one of the other agent in the turn. They are aware of the complete sentence and premise pool.

After argument introduction, all agents recognise whether they need to update their position or not. Those that do are aware of all of their options, i.e. they know all the remaining coherent positions in the SCCP and their own position's distance to them.

Interaction: Agents interact in terms of argument introduction described above. Agents are indirectly influenced by the actions of other agents when their position is rendered incoherent and they are forced to update.

The interaction of agents is affected only by random processes. For example, agents do not prefer to introduce attack arguments against agents with a high Hamming distance to their position. They are also ignorant to what relations their introduced argument will have to existing arguments.

Agents impact the opinion dynamics of others by introducing logical constraints to the debate. For a minimal example, consider two agents with positions $a_1 = \{p_1 \rightarrow \text{True}, p_2 \rightarrow \text{True}, p_3 \rightarrow \text{False}\}$ and $a_2 = \{p_1 \rightarrow \text{False}, p_2 \rightarrow \text{False}, p_3 \rightarrow \text{True}\}$, and consider that a_1 would introduce the valid argument $(p_1 \wedge p_2) \Rightarrow \neg p_3$, thus reflecting its truth-value attributions. The argument stands against a_2's belief in p_3, but is a_2 forced to

update its system of belief because of the argument? No. a_2 does not accept p_1 and p_2, and so does not need be moved by an argument that relies on their truth. If a_2 would accept both p_1 and p_2, then giving up p_3 would result in the shift to the closest coherent position.

This simple example illustrates how agents only have intermediate control over the beliefs of others. Through their argument introductions, agents shape the space of complete and coherent positions. But what the other agents make of their options is a different issue.

Collectives: Agents are not grouped into collectives. The population is regarded as a uniform whole.

Observation: Among others, the model tracks every agent and its position at every debate stage and the density of that stage. These are the two fundamental variables for calculating polarisation measures. The current implementation of the model logs more information about the model run, including position updating at each stage, and all of the arguments introduced at any given debate stage.

A.3 Details

A.3.1 Initialisation

At the start of every simulation, the sentence pool is generated. In the main simulation presented below, the sentence pool is generated from 20 atomic sentence variables, $p_1, p_2, ..., p_{20}$, and their negations, $\neg p_1, \neg p_2, ..., \neg p_{20}$. The sentence pool thus consists of 40 sentences. From this pool, a premise pool is constructed. The premise pool consists of all combinations of sentences that can be used in an argument. Given that the argument length is set to 2 or 3 premises and given the condition that an atomic sentence variable should appear only once in the premises of each argument, the number of possible combinations of premises is 9880:

$$\binom{40}{2} - 20 + \binom{40}{3} - 20 \cdot 38 = 760 + 9120 = 9880$$

Agents select their initial position by randomly assigning truth values to every atomic sentence variable (though see the robustness analysis in §5 which varies this initial setting). In the simulations below, these truth values are either True or False, and simulating positions with probabilistic assignments is left for future research. The debate contains no arguments at the beginning of the simulation.

What kind of debates are simulated with the initial values for the main experiment? With 50 participants and 40 sentences under discussion, these artificial debates could model political deliberation in parliament or scientific deliberation at a conference (think of a panel and its audience), or of the participants at a citizen deliberation event. However, the fact that there is a continuous

debate forum and no side conversations take place is a simplification over the real-world originals, while the restriction of arguments with 2 and 3 premises is due to computational limitations.

A.3.2 Input

Input to the model is limited to the settings in the model initialisation. Environment variables such as the premise pool and the space of coherent and complete positions change only based on the agents' behaviour, as described in the two sub-modules.

A.3.3 Sub-modules

There are two noteworthy sub-processes that shape the evolution of debates in the model. Both have elements of random choice – which is why the model should be evaluated in simulation experiments with many runs.

Argument introduction: At every debate stage, two agents from the population are drawn at random. Depending on the argumentation strategy, the first agent (the source) then draws premises from the premise pool that meet the criteria imposed by the strategy. For example, in the convert strategy it will draw a random set of premises that is accepted by the target agent. The fortify strategy is a special case in this regard, since it does not require inspecting the target's belief system.

Next, the source agent looks for a conclusion from the sentence pool that (a) is not equivalent or contradictory to one of the premises and (b) meets the criteria that the argumentation strategy imposes on conclusion choice. For example, in the convert strategy, this conclusion must be one that is currently accepted by the source agent. The search continues until a valid argument is found, i.e. one that is jointly satisfiable in conjunction with the arguments already present in the debate. When the introduction succeeds in this manner, the set of selected premises is removed from the premise pool, which means that this set of premises is unavailable for subsequent argument introductions for the rest of the model run. If no valid argument can be found for this particular pair of agents, the process is repeated at most $A/2$ times by drawing another pair of agents from the population, where A is the size of the population.

It should be noted that argument introduction almost always changes the extension of the space of coherent and complete positions, although argument introductions can differ significantly in their impact. Argument introductions can render previous positions incoherent, and are the driver behind agents updating their positions in the course of the debate.

Position updating: Newly introduced arguments can render existing positions incoherent, and they regularly do. After each argument introduction, *all*

agents in the debate check whether their position is still valid given the new debate stage, and all agents that now hold incoherent positions update them.

For all agents, the update strategy in this model is always the move to the closest coherent position. In order to find the closest coherent, every agent with an incoherent position compares its position to all coherent positions in the SCCP, and moves to the one with minimal Hamming distance. If there are multiple positions with minimum distance, one is chosen at random.

Appendix B

The user guide to taupy, a Python package to study the theory of dialectical structures

This is the user guide for `taupy`, a Python package for the study of dialectical structures. The name for the package derives from the Greek letter τ ("tau"). This letter is used in the theory of dialectical structures to describe its fundamental object. `taupy` was designed and implemented as part of this PhD thesis. The source code is published as Kopecky (2021–2022) and is available on GitHub at https://github.com/kopeckyf/taupy.

The user guide is intended for new users and developers of `taupy`, to be read alongside the source code. It was slightly adapted for this PhD thesis from the original documentation at https://kopeckyf.github.io/taupy/. Some readers might find this web version more accessible.

B.1 Installation and quick start

B.1.1 Installation

`taupy` is distributed on PyPi, the Python Package Index. The project page for `taupy` is available at https://pypi.org/project/taupy/. The latest release version can be installed with the following console command:

```
> pip install taupy
```

To install `taupy` from source, clone the GitHub repository from https://github.com/kopeckyf/taupy and run

```
> pip install ./
```

in the root folder containing `setup.py`.

B.1.2 Quick start

After installation, `taupy` can be imported to Python either by importing the module to the current name space:

```
import taupy
```

or by importing every public object from `taupy` to the current name space:

```
from taupy import *
```

> **Note**
>
> All objects in the user guide can be publicly accessed even though they are often referenced with their full path. For example, the class `taupy.basic.core.Argument` can be accessed simply as `Argument` if all `taupy` objects have been imported to the current name space using the `*`-notation, or as `taupy.Argument` if the module has been imported to the current name space. It would also be possible to use the pedestrian reference `taupy.basic.core.Argument` when the module has been imported.

B.1.3 Known installation issues

There are some known problems during the installation of dependencies that `taupy` relies on. These can usually be solved by installing these packages manually before initiating the `taupy` installation in pip.

B.1.3.1 On Windows

Prior to version 0.5, `taupy` relied on `iteration-utilities`. This package implements functions in pure C, and its source code must be compiled before it can be used in Python. Unfortunately the package became unmaintained around Python 3.10 and no pre-compiled wheel packages were available. Windows users who use Python 3.10 or newer and `taupy` 0.4 or older are particularly affected by this, as they need to install Microsoft's Visual C++ Build Tools in order to compile the C code.

As of version 0.5, `taupy` no longer depends on `iteration-utilities` but uses `more-itertools` for advanced combinatorial tasks. `taupy` version 0.4 or older can be used with Python 3.9 on Windows without installing additional build tools.

B.1.3.2 On Mac OS

`taupy` relies on the `scipy` Python package for scientific computing. No wheels are provided for `scipy` on the ARM version of Mac OS 11 via pip. On ARM Macs ("Apple Silicon") that run Mac OS 11, `scipy` needs to be pre-installed via conda before `taupy` can be installed via pip. This issue can also be resolved by upgrading to a newer version of Mac OS.

B.1.3.3 On Linux

No installation issues are known on any Linux distribution.

B.2 Base objects

The theory of dialectical structures is an abstract representation of deliberation. It uses three main object types to do so. A `Debate` is composed of a set of arguments as well as a support and defeat relation between them. An `Argument` consists of premises and a conclusion. The relations between arguments are automatically determined through their premises and conclusions. The belief systems held by agents in light of the debate are described as a `Position`.

B.2.1 Arguments

TDS understands arguments as premise-conclusion structures. This means that any argument has a set of premises $\{a_1, a_2, ..., a_n\}$, a rule of inference r and a conclusion c. `taupy` does not implement the rule of inference since it assumes logical validity for any argument. Then, the logical structure of an argument is that of an implication: $(a_1 \wedge a_2 \wedge ... \wedge a_n) \Rightarrow c$.

`taupy` implements arguments as implication relations from `sympy`, a package for symbolic computing in Python. `taupy` entirely relies on `sympy` for symbolic manipulation and some Boolean algebra. In other words, the **`taupy.basic.core.Argument`** class is a sub-class of `sympy.Implies`.

For the creation of `Argument` instances, sentence variables need to be present. In the interactive mode, it is recommended to create such objects as `sympy.Symbol` objects via the `sympy.symbols` function. It is impossible to use sentence variables without declaring them first. This is due to a core principle in Python: every variable needs to be declared before it can be used.

```
from taupy import Argument
from sympy import symbols

a, b, c = symbols("a b c")
# Alternatively, but for limited amount of variables only:
# from sympy.abc import a, b, c, d, e, ...
```

Now that we have three sentence variables, we can construct a simple argument `a1` with two premises and one conclusion:

```
a1 = Argument(a&b, ~c)
```

The premises `a&b` are conntected by the operator &, not by a comma. Premises and conclusion make up two parameters for the instance of an `Argument`, and these are separated by a comma. We have taken the negation of `c` as a premise here by using the `sympy` operator ~. Alternatively, we could have used

`sympy.Not`. We could have done the same for any of the premises. And we could have entered further premises by adding them with a `&`.

B.2.2 Debates

A debate in the theory of dialectical structures is a set of arguments on which two relations are defined: support and defeat. Formally, a debate is a tuple $\tau = \langle T, A, U \rangle$, where T is the set of arguments, A the pairs of arguments that fulfil the defeat relation, and U the pairs of arguments that fulfil the support relation.

B.2.2.1 Relations between arguments in a debate

The defeat and support relation in TDS are defined as follows: An argument $a \in T$ defeats another argument $b \in T$ if the conclusion of a is equivalent to the negation of a premise in b. An argument $a \in T$ supports another $b \in T$ if the conclusion of a is equivalent to one of the premises in b.

`taupy` does not require manual input of relations. They are determined automatically from the provided arguments. See Figure B.1 for an example with ten arguments and some defeat and support relations.

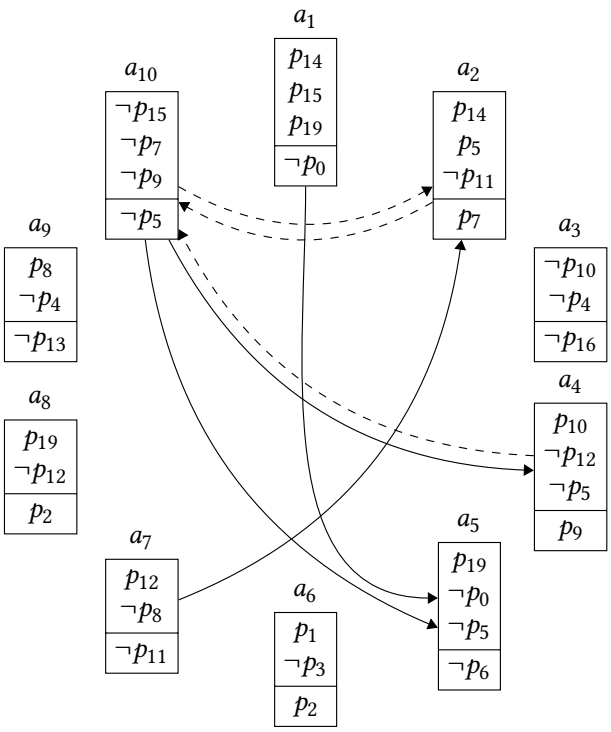


Figure B.1: An example for a dialectical structure with 10 arguments. Some arguments defeat or support others. Supports between arguments are visualised with solid edges, and defeats through dashed edges.

B.2.2.2 Debate objects

In `taupy`, debates are instances of the `taupy.Debate` class. The general logical structure of a debate is that of a conjunction: **`taupy.basic.core.Debate`** is a sub-class of `sympy.And`.

A `Debate` is composed of arguments. These do not necessarily exhibit any relation. When a debate is initialised, its arguments are given as a comma-separated list. Any number of arguments can be passed to a `Debate` object.

```
from taupy import Argument, Debate
from sympy import symbols
a, b, c, d, e = symbols("a b c d e")
tau1 = Debate(Argument(a&b, ~c), Argument(~d&e, ~a))
```

`tau1` is a debate with two arguments. The second `Argument(~d&e, ~a)` defeats the first `Argument(a&b, ~c)`.

B.2.3 Positions

B.2.3.1 Position objects

(Valid) arguments are relations between propositions: if the premises were true, the conclusion would be as well. Agents can accept arguments as valid without agreeing on the premises. The mere existence of an argument does not force an agent to assert the conclusion if it does not accept one or more of the premises.

The theory of dialectical structures represents belief systems of agents as mappings from the propositions discussed in a debate to truth values – positions can only be held relative to a debate. A position that assigns True to a proposition is said to "accept" it, and to "reject" it if it assigns False.

Let us look at a debate τ_1 with five propositions: a, b, c, d and e:

$$\tau_1 = ((a \wedge b) \Rightarrow \neg c) \wedge ((\neg d \wedge e) \Rightarrow \neg a)$$

and a position Pos_1 that assigns truth values to these propositions.

$$\mathrm{Pos}_1 = \left\{ \begin{array}{l} a \rightarrow \mathrm{True} \\ b \rightarrow \mathrm{False} \\ c \rightarrow \mathrm{True} \\ d \rightarrow \mathrm{True} \\ e \rightarrow \mathrm{False} \end{array} \right\}$$

In `taupy`, positions are input as instances of **`taupy.basic.positions.Position`**:

`class taupy.basic.positions.Position(debate, *args, introduction_strategy=None, update_strategy=None)`

As positions can only be held relative to a debate, it is given as the first argument when creating a new `Position` object. The truth-value assignments are given as a mapping in the second argument:

```
from taupy import Argument, Debate, Position
from sympy import symbols
a, b, c, d, e = symbols("a b c d e")
tau1 = Debate(Argument(a&b, ~c), Argument(~d&e, ~a))
pos1 = Position(tau1, {a: True, b: False, c: True, d: True,
                       e: False})
```

B.2.3.2 Properties of positions

Completeness Positions do not necessarily assign a truth value to every sentence in the pool. A position that does not is called "partial". A "complete" position assigns a truth value to every proposition in the debate.

```
pos1 = Position(tau1, {a: True, b: False, c: True})
# Check whether it is complete, i.e. assigns True or False
# to every sentence in its Debate:
pos1.is_complete()
```

Closedness A position is closed if it follows its dialectical obgliations: if a position assigns True to all premises in an argument, it must also assign True to the conclusion. Check this property using `Position.is_closed()`.

`taupy.basic.positions.closedness(pos, debate=None, return_alternative=False)`

A position `pos` is closed relative to a debate when it follows its dialectical obligations: if a position assigns True to all of the premises of an argument in the debate, it must also assign True to the conclusion of that argument.

This function assumes that the input `pos` is coherent. If in doubt, you should perform a coherence check first. Incoherent positions can be labelled as closed by this algorithm, although this is nonsensical.

Returns a Boolean by default indicating the closedness status of `pos`. However, if `return_alternative` is True, the function will return a tuple containing the closedness value and an alternative. The alternative is obtained by checking if the position follows entailment. If the position is closed, the alternative will be the position itself, but in case of closeness violation, the function will close the position by filling up the position via entailment.

A shortcut of this function exists under `Position.is_closed()`.

Coherence There are two ways to express coherence for a position relative to a debate:

1. The position (a) assigns identical truth values to sentences that are equivalent give the debate and complementary truth values to incompatible ones and (b) does not contradict its inferential obligations, i.e. if it accepts all premises of an argument, it can not reject the conclusion.

2. The position satisfies the Boolean formula that represents the debate. In other words, if SAT(τ) returns the set of satisfying assignments of the Boolean formula for the debate τ, then a position p is coherent just in case $p \in \text{SAT}(\tau)$.

> **Note**
>
> Coherence does not imply closedness. Under condition (1b), a coherent position is only required not to contradict its inferential obligations. But a coherent position can still not follow some of them, e.g. by being a partial position and not assigning a value to the conclusion. Coherence and completeness jointly imply closedness.

```
pos1 = Position(Debate(Argument(a&~b, c)),
                {a: True, b: False, c: False})
# Will return False:
pos1.is_coherent()
```

B.2.3.3 The set of coherent and complete positions

A central concept to the theory of dialectical structures is the collection of all positions that are coherent and complete (and thus closed) given a debate.

Given a debate `tau`, this set can be obtained with `taupy.satisfiability(tau, all_models=true)`.

`taupy.basic.utilities.satisfiability(f, all_models=True)`
Return a generator of models for the given Boolean formula `f`, using BDDs.

When interpreted as a graph, this set is called the *space of coherent and complete positions*, or SCCP, often expressed by the Greek upper-case letter gamma (Γ). The SCCP yields insights about the debate itself, such as its inferential density.

Debate.sccp(return_attributions=False)

Returns a dictionary of lists

```
{position: [neighbour1, neighbour2, …], …}
```

that resembles the space of coherent and complete positions. This structure serves as the basis for graph analysis and graph drawing.

Iteration is done over the possible neighbours of a position rather than with all other positions, because the searches' complexity will be lower.

If `return_attributions` is set to True, this function returns a tuple. The first object then is the graph representation in a dict of lists format, the second object is a mapping from the string representation of a position to its dictionary format. This is useful because non-hashable objects like dictionaries can not be used as identifiers of nodes in graphs.

```
tau1 = Debate(Argument(a&~b, c))
tau1.sccp()
```

B.3 Analysis & measurement

B.3.1 Agreement and distance

The distance between two positions is always based on their differences in truth-value assignments, and their agreement is closely related to this difference: if δ is a normalised distance function, then the normalised agreement of two positions x and y is given as $1 - \delta(x, y)$.

The distance functions below accept a pair of agents and output the agreement or distance of that pair. To obtain the differences among more than two positions, use `difference_matrix` to obtain a square matrix of distances.

taupy.analysis.polarisation.difference_matrix(positions, measure)

Create a quadratic matrix D_{ij} in which rows and columns are filled by `positions`. The value at d_{ij} is the distance, calculated by `measure`, between positions i and j.

This matrix of distances is the fundamental object to calculate most polarisation measures.

B.3.1.1 Hamming distance

For positions that assign truth values to the same propositions, particularly for complete positions of the same debate, the Hamming distance is the most easy distance measure. It simply counts the items that two positions evaluate differently.

`taupy.analysis.agreement.hamming_distance(pos1, pos2)`
Returns the Hamming distance between two positions of equal length. The Hamming distance counts the number of differences in truth-value attributions. This distance can only be calculated for positions with the same domain (complete positions on the same debate have the same domain).

The Hamming distance can be normalised to the number of proposition to which the positions assign truth values (their "length").

`taupy.analysis.agreement.normalised_hamming_distance(pos1, pos2)`
Returns the Hamming distance between `post1` and `pos2`, normalised by the number of propositions in the positions' domain.

Closely related to the Hamming distance is Betz's normalised agreement, or `bna()` for short. For two positions x and y of equal domain,

$$\mathrm{bna}(x, y) = 1 - \mathrm{HD}(x, y)/\mathrm{len}(x).$$

`taupy.analysis.agreement.bna(pos1, pos2)`
A normalised agreement measure for positions of equal length, which is used by Betz (2013, p. 39). Here, agreement is normalised to the length of the positions.

B.3.1.2 Edit distance

The notion of difference and agreement is meaningful for positions of different domains as well. The edit distance is equal to the minimal number of operations that are necessary to transform one position into the other. Each operation is assigned to a weight, and the edit distance is calculated as a weighted sum of the operations. There are three operations: switching of a truth value, adding of a truth-value assignment, and deletion of one. The edit distance is generally asymmetric if the weights are unequal.

> **Hint**
>
> When all operations in the edit distance have the same weight, the edit distance is a generalisation of the Hamming distance. For positions of equal domain, it then simplifies to the Hamming distance.

`taupy.analysis.agreement.edit_distance(pos1, pos2, weights = {'deletion': 1.0, 'insertion': 1.0, 'substitution': 1.0})`
A generalised distance measure that does not require that the positions share their domain. Compared to edit distances for ordered sequences

(e.g. Levenshtein distance), it is far easier to compare two positions in terms of TDS.

For each item, two positions can have four states:

Agreement: They agree on the item, which does not increase the distance.

There are three operations that do increase the distance:

Substitution: The positions are equal after transposition, i.e. changing a truth-value

Insertion: One position does not make a statement concerning one proposition. Adding the respective truth-value attribution makes the two positions equal w.r.t the statement.

Deletion: One position does make a statement that the other does not care about. The positions can be made equal if the first position forgets its statement.

The edit distance can be normalised by first calculating the maximal distance given the union of the positions' domains and the weights allocated to the operations.

`taupy.analysis.agreement.normalised_edit_distance(pos1, pos2, weights = {'deletion': 1.0, 'insertion': 1.0, 'substitution': 1.0})`
The (weighted) edit distance, normalised to return a value in $[0, 1]$. Normalisation is understood as the relation between actual and maximal difference. Maximal difference is achieved in the edit distance if the most costly action is performed for all items.

`taupy.analysis.agreement.normalised_edit_agreement(pos1, pos2)`
An agreement function based on the normalised edit distance is defined for convenience. It equals $1 - \mathrm{ED}_n(x, y)$.

B.3.2 Inferential density

Inferential density measures how much the arguments in a debate have constrained the space of coherent positions, which indicates how free agents are in their choice of an admissible belief system in light of the debate.

Consider how the debate consisting of just one argument, $(a \wedge b) \Rightarrow c$ has $2^3 - 1$ coherent and complete truth-value assignments. The one that is missing assigns False to c but True to a and b. Betz (2013, p. 44) gives a general formula for calculating this density of a debate τ, $D(\tau)$, with a sentene pool of length n and a space of complete on coherent positions $\Gamma(\tau)$:

$$D(\tau) := \frac{n - \log_2(|\Gamma(\tau)|)}{n}$$

Density offers a measure independent of simulation time that captures the progress in debates. It is a reliable alternative to the number of introduced arguments. A debate in which the arguments impose rather few constraints on the available complete and coherent positions will have a low density; a debate in which this influence is high, the debate's density will be high as well. However, while the density generally rises with the number of arguments, not every argument renders a previously coherent position incoherent. And so, not every argument contributes to density equally, and some won't change it at all.

`taupy.basic.core.Base.density()`

Return the dialectical density of the Debate object, as defined by Betz (2013, pp. 44–49).

Density is returned as a fraction. A floating number can be obtained from the result using Python's `float()`.

Hint

`density()` is implemented as a method to the `Base` object. As both `Argument` and `Debate` inherit from `Base`, density is available as a method for these objects, too.

B.3.3 Degrees of justification

A degree of justification (DOJ) quantifies how justified a truth-value assignment is in light of a debate. The concept was introduced to the theory of dialectical structures in Betz (2012). A DOJ always lies in the interval $[0, 1]$ – and can be treated as a probability in the sense that it fulfils the Kolmogorov axioms (Betz 2012, Theorem 6).

`taupy.analysis.doj.doj(pos, debate=None, conditional=None)`

Returns the degree of justification for the position in `pos` relative to a `debate`.

If `debate` is `None`, the debate stored in the Position object `pos` is used.

The *conditional* doj is returned if `conditional` is given another position of the same debate. When `conditional` is set, `debate` must be `None`.

Hint

The fraction output from `taupy.doj` can be converted to float using `float(doj())`.

B.3.3.1 Unconditional DOJs

Let P be a (partial) position in the debate τ and $\Gamma(\tau)$ the space of coherent and complete positions on τ. Then, the degree of justification of P in τ is defined as follows.

$$\mathrm{doj}(P)_\tau := \frac{|\{\gamma \in \Gamma_\tau | P \subseteq \gamma\}|}{|\Gamma_\tau|}$$

The DOJ of a position P in a debate is equal to the proportion of positions in the debate's SCCP that extend P, or have its truth-value assignments as a part. One can also understand this in terms of probability: if all complete and coherent positions in a debate τ were equally likely of being drawn, then how likely would the set of propositions P be true according the drawn position?

```
from taupy import Argument, Debate, doj
from sympy.abc import a, b, c
# returns 3/7
doj({c: False}, debate=Debate(Argument(a&b, c)))
```

> **Note**
>
> The DOJ of an incoherent position always equals zero. The DOJ of a complete position equals $1/|\Gamma_\tau|$, since there is exactly one item in the SCCP that extents that position – and this is the position itself. This means that DOJs are most informative for coherent partial positions, such as truth-value assignments of single sentences.

B.3.3.2 Conditional DOJs

We can not only ask the question of how well a position is justified given a debate simpliciter, but also how well it would be justified if some statements in the debate were taken for granted. Let C be a set of propositions relative to which the justification of P should be evaluated.

$$\mathrm{doj}(P|C)_\tau := \frac{|\{\gamma \in \Gamma_\tau \cap C | P \subseteq \gamma\}|}{|\{\gamma \in \Gamma_\tau \cap C\}|}$$

In `taupy`, a conditional DOJ is calculated with the `conditional` argument:

```
from taupy import Argument, Debate, Position, doj
from sympy.abc import a, b, c
pos1 = Position(Debate(Argument(a&b, c)), {a: True})
pos2 = Position(Debate(Argument(a&b, c)), {c: True})
# What is the degree of justification for pos1,
# conditional to pos2? (It's 1/2)
doj(pos1, conditional=pos2)
```

B.3.4 Centrality

A population of positions can be interpreted as a graph in which the individual agents make up the nodes of the graph, and the edge weights are determined by the distance between their belief systems. Agents can inhibit the centre of a graph or be far removed at the fringes. The normalised closeness centrality indicates this position for a single agent in a population relative to a distance measure (see Betz 2013, Section 2.4). In Figure B.2, the top position has a distance of 1 to each of the others, and the others have a higher distance towards each other. The top agent thus is the most central node.

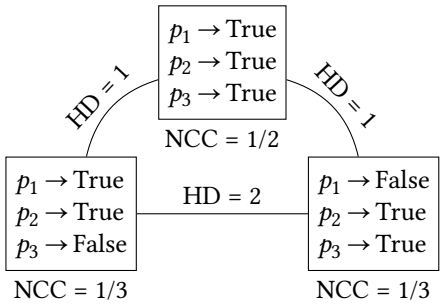


Figure B.2: Three positions, their mutual distances and normalised closedness centrality scores.

Note

There can be multiple, far removed agents with maximum NCC in a population. In other words, a population graph can have multiple graph centres.

`taupy.analysis.agreement.ncc(population, *, agent, measure = <function hamming_distance>)`

Returns the normalised closedness centrality (NCC) of an `agent` in a `population` relative to a `measure`.

Parameters:

`population:` Iterable containing agents' belief systems.

`agent:` A single belief system.

`measure:` A distance measure between belief systems.

B.3.5 Clustering belief systems

Clustering is a pre-condition for all diversity and some polarisation measures. Intuitively, clusters should be formed by groups that have low internal distance. Belief systems in the theory of dialectical structures are clustered by

comparing their distances to other agents. Clustering depends on a distance function in this way.

B.3.5.1 Generate clustering matrices

The clustering algorithms accept a `clustering_matrix` as input, which is generated by `clustering_matrix()`:

`taupy.analysis.clustering.clustering_matrix(positions, *, measure = <function normalised_hamming_distance>, scale = -4, distance_threshold = 0.2)`

Converts a difference matrix to a sparse clustering (adjacency) matrix that can be input to community structuring algorithms. This is necessary because many clustering algorithms are designed for sparse social networks.

The default scale of -4 means that agents with a normalised $\delta > 0.4$ will not be treated as adjacent to each other.

B.3.5.2 Clustering algorithms

taupy implements four clustering functions. These functions serve as front-ends to the clustering algorithms implemented in `igraph` (Csárdi & Nepusz 2006) and `sklearn` (Pedregosa et al. 2011).

`taupy.analysis.clustering.leiden(positions, *, clustering_settings={})`

Return the community structure obtained by the Leiden clustering algorithm (see Traag, Waltman & van Eck 2019).

`taupy.analysis.clustering.affinity_propagation(positions, *, clustering_settings={})`

Return the community structure obtained by clustering with Affinity Propagation (Frey & Dueck 2007).

`taupy.analysis.clustering.agglomerative_clustering(positions, *, distance_threshold=0.75, base_measure=<function normalised_hamming_distance>)`

Return community structuring obtained by Agglomerative Clustering. Please note that Agglomerative Clustering accepts a common difference matrix, not an adjacency matrix as Leiden and Affinity Propagation do. It is not advisable to pass the output of `clustering_matrix()` to this function. Please use `difference_matrix()` with a normalised distance measure as input.

`taupy.analysis.clustering.density_based_clustering(positions, *, min_cluster_size=3, max_neighbour_distance=0.2, base_measure=<function normalised_hamming_distance>)`

Return community structure obtained from density based clustering on a distance (not adjacency) matrix. This clustering algorithm is the only one implemented in this module to allow noise. Points with -1 signal noise.

B.3.5.3 Comparing clusterings

The adjusted Rand index (ARI, Hubert & Arabie 1985) measures the similarity between two clusterings. For simulations that contain many debate stages, the ARI can indicate whether the clustering in subsequent debate stages has completely changed or is somewhat stable. A reasonably high ARI can support the reliability of the clustering method for the analysed debate stages.

`taupy.basic.utilities.ari(partition1, partition2)`
Calculate the Adjusted Rand Index.

B.3.6 Polarisation

Identifying a population as polarised depends on a measure of polarisation as there are many different notions of "polarisation" (Bramson et al. 2016). `taupy` implements many of the measures identified by Bramson et al., but in a definition that is adapted to the belief systems in the theory of dialectical structures.

B.3.6.1 Measures without clustering

`taupy.analysis.polarisation.spread(positions, measure)`
Returns the maximum distance between any two of the positions relative to a measure (originally defined in Bramson et al. 2016, pp. 80–111).

`taupy.analysis.polarisation.pairwise_dispersion(positions, measure)`
Returns dispersion, understood as the standard deviation of pairwise distances between the `positions` relative to the `measure`. This measure was defined in Bramson et al. (2016).

This is the TDS equivalent of statistical dispersion or variance in polling data. Beside standard deviation, there are other ways of measuring dispersion.

Bramson et al. take the dispersion relative to a mean. However, since such a mean is not well-defined in TDS, we use dispersion on pairwise relations instead.

For this purpose, we use the upper triangle of the difference matrix, without the diagonal zeroes (this offset is controlled by $k = 1$). Since $D_{a,b} = D_{b,a}$, these are the pairwise difference values we are after. We then take the standard deviation of these values.

B.3.6.2 Measures with clustering

taupy.analysis.polarisation.group_divergence(clusters, adjacency_matrix)

A variant of Bramson et al.'s group divergence (Bramson et al. 2016), adapted to belief systems in the theory of dialectical structures.

Group divergence relies on a useful clustering that returns `clusters`, which is expected to be a list of lists. The `adjacency_matrix` can be a modified or scaled version of `difference_matrix()`, or the verbatim matrix.

Algorithms which are known to being able to return good results in TDS are:

- Leiden (implementation from `python-igraph`) and other modularity maximisation approaches.
- Affinity propagation (implementation from `scikit-learn`).
- Agglomerative clustering (implementation from `scikit-learn`).

taupy.analysis.polarisation.group_consensus(clusters, adjacency_matrix)

A variant of Bramson et al.'s measure of group consensus (Bramson et al. 2016), adapted to belief systems in the theory of dialectical structures.

As `group_divergence()`, this relies on a good clustering as well. Arguments and recommendations for algorithms to try are the same as in `group_divergence()`.

taupy.analysis.polarisation.group_size_parity(clusters)

Bramson et al.'s (2016) measure of (group) size parity, adjusted to belief systems in the theory of dialectical structures. According to the authors, size parity is an entropy measure, which is irrespective of the size of the population and of the number of groups. It is said to behave erratically in case the groups are determined endogenously, e.g. by one of the clustering algorithms Leiden, Affinity propagation, etc.

taupy.analysis.polarisation.coverage_of_clustering(clusters, *, noise_value=-1)

The amount of agents fitted into a cluster by algorithms with noise. Examples for such algorithms are `OPTICS` and `DBSCAN` from sklearn.

Algorithms that do not allow noise will always return a coverage of 1.

Non-clustered nodes are marked by -1 by convention.

B.3.7 Diversity

Functions to measure diversity among deliberating agents. Many functions in the module have a common abstract ancestor function (Tuomisto 2010), but they are implemented here in a rather straightforward way for simplicity.

B.3.7.1 Shannon index and derivatives

`taupy.analysis.diversity.Shannon_index(clusters)`
: The Shannon index describes the uncertainty of predicting the cluster that the belief system of a randomly drawn agent belongs to.

`taupy.analysis.diversity.normalised_Shannon_index(clusters)`

B.3.7.2 Simpson index and derivatives

`taupy.analysis.diversity.Simpson_index(clusters)`
: The Simpson index of diversity equals the probability that the belief systems of two randomly chosen agents belong to the *same* cluster.

`taupy.analysis.diversity.inverse_Simpson_index(clusters)`
: Simpson's inverse index is simply dubbed "diversity index" by Page (2011, pp. 73–76). Political scientists call it "effective number of parties", and it is known as Herfindahl index in economics.

`taupy.analysis.diversity.Gini_Simpson_index(clusters)`
: Estimates the probability that the belief systems of two randomly drawn agents are clustered into *different* clusters ("inter-type encounters").

B.3.7.3 Attribute diversity

`taupy.analysis.diversity.attribute_diversity_page(positions)`
: This function is named after Page's (2011, pp. 73–76) "attribute diversity". This diversity index is equal to the number of distinct attributes in the population. We interpret it to count the number of distinct truth-value attributions: a population in which both $\{p1 \rightarrow \text{True}\}$ and $\{p1 \rightarrow \text{False}\}$ are maintained is more diverse then a population in which just $\{p1 \rightarrow \text{True}\}$ is maintained.

  Note that this is not quite the same as *richness*, which would be equal to the number of clusters.

`taupy.analysis.diversity.normalised_attribute_diversity_page(positions, sentencepool, truth_values=[True, False])`
: Page's attribute diversity, normalised to the amount of truth-value attributions possible without any constraints (number of sentences * allowed truth values). This normalised diversity measure is not weighted, since all attributes contribute to it equally.

B.4 Synthesise new objects

This module provides tools for generating debates and positions from more basic objects. Examples are hierarchical argument maps, in which arguments

are arranged to form a tree-like graph, and methods for aggregating the beliefs of many agents into a single belief system.

B.4.1 Surveys and aggregation

`taupy.analysis.voting.survey(p, *, positions, not_present_value=None)`
Take a survey about proposition `p` among `positions`. This function is agnostic about truth values attributed to `p`. These could be True and False assuming a two-valued logic. If any position does not pass a judgement on `p`, they respond `not_present_value` to the survey.

`taupy.analysis.voting.majority_vote_winner(p, *, positions, not_present_value=None)`
Cast a simple majority vote about `p` among `positions` and return the winner. This function checks whether the winner would be unique. A value error is returned if no winner can be determined due to a tie.

`taupy.analysis.voting.aggregated_position_of_winners(positions, *, not_present_value=None)`
Return a `Position` that is aggregated from the input positions by majority voting.

B.4.2 Argument maps

`taupy.generators.maps.generate_hierarchical_argument_map(N = 20, k=3, max_num_args=inf, max_density=1.0, distribution={2: 0.19, 3: 0.23, 4: 0.32, 5: 0.26}, base_conclusion=0.75, base_premises=0.75)`
Generate a hierarchical synthetic argument map, following the algorithm from Betz, Chekan & Mchedlidze (2021).

B.5 Simulations

`Debates` and `Positions` are static objects: they describe arguments and an agent's belief system at one point in time. Dynamical aspects of debates and belief systems can be studied in `Simulations`.

In agent-based models built on the theory of dialectical structures, the simulated world consists of a sentence pool and arguments made from these sentences. Agents have a multi-dimensional belief system in terms of `Positions`. The `Simulations` progress by introducing arguments, possibly according to an argumentation strategy assigned to agents. Arguments can be introduced in a random fashion to debates. Those have minimal requirements and can be used even in Simulations that do not contain any agents. Purposeful argumentation strategies select premises and conclusions in light of agents' belief systems. For these arguments, at least two agents are drawn from the population.

After argument introduction, the entire population responds by checking their belief system in light of the new argument and revise their beliefs if necessary. This process continues until a termination condition is reached. If requested by the user, the sentence pool is also occasionally extended. The simulation terminates when the desired inferential density is reached or the desired number of arguments has been introduced.

B.5.1 Setting up a population

Simulations in which agents update their belief systems need to be initialised with a list of positions. This is optional: a debate can progress without any agents present.

When simulations are initialised with a population, the agents from this population update their belief systems in response to argument introductions and sentence pool expansions. The simulation objects have a `Simulation.positions` attribute, which is a list of populations in which the *i*th element stores the population at the *i*th simulation step. The initialised population is stored in the first element, `Simulation.positions[0]`.

B.5.1.1 Populations are lists of positions

The initial population needs to be generated as a list of taupy.Position objects:

```
my_population = [Position() for _ in range(10)]
```

The agents' behaviour in argument introductions can be controlled using the `introduction_strategy` attribute.

```
my_population = [Position(debate=None,
    introduction_strategy=strategies.fortify \
    for _ in range(10)]
```

A population can consist of agents with different argumentation strategies (see the `taupy.simulation.strategies` module for pre-defined strategies. Custom argumentation strategies are also accepted):

```
fortify_positions = [Position(debate=None,
    introduction_strategy=strategies.fortify \
    for _ in range(5)]

convert_positions = [Position(debate=None,
    introduction_strategy=strategies.convert \
    for _ in range(5)]

my_population = fortify_positions + convert_positions
```

When you initialise positions like this, they will be assigned random truth-value attributions during the simulation initialisation.

It is also possible to generate a population with specific beliefs. The population below is set-up as bi-polarised:

```
pos_template_1 = {
    symbols("p0"): True, symbols("p1"): True,
    symbols("p2"): True, symbols("p3"): True,
    symbols("p4"): True, symbols("p5"): True
}

pos_template_2 = {
    symbols("p0"): False, symbols("p1"): False,
    symbols("p2"): False, symbols("p3"): False,
    symbols("p4"): False, symbols("p5"): False
}

pop_part_1 = [Position(None, pos_template_1,
              introduction_strategy=strategies.convert) \
              for _ in range(10)]
pop_part_2 = [Position(None, pos_template_2,
              introduction_strategy=strategies.undercut) \
              for _ in range(10)]

my_polarised_pop = pop_part_1 + pop_part_2
```

B.5.1.2 Argumentation strategies

The following argumentation strategies are pre-defined in taupy. They all introduce arguments that are valid given the current debate stage.

`random` A completely random strategy that works even if a simulation has no positions at all.

`fortify` Insert a valid argument the premises and conclusion of which are accepted by the source position.

`attack` A valid argument. The premises are accepted by the source position, and the source at least tolerates the conclusion. However, the target denies the conclusion, given its current truth-value attribution.

`convert` A valid argument with premises picked from the target. The conclusion is picked from the source, and the source also accepts the conclusion. It is not checked whether the target accepts the conclusion.

`undercut` A valid argument is constructed with premises that the target accepts. The source at least tolerates the conclusion. The conclusion however is not accepted by the target.

`unrestricted_undercut` Like undercut, but does not require the source agent to tolerate the conclusion.

B.5.2 Setting up a simulation

Simulations are instances of a simulation class. There are multiple simulation classes in `taupy` for different kinds of simulations.

- The class `taupy.Simulation` composes arguments at each introduction step.
- In the class `taupy.FixedDebateSimulation`, argument maps are precompiled and arguments are individually uncovered in each introduction step.

B.5.2.1 Examples

A minimal example Simulations without agents that follow purposeful introduction or updating strategies can be created with a call to `Simulation` and leaving the `positions` to `None`. Simulations without positions can only contain un-directed, random arguments. Introduced arguments can only follow a purposeful argumentation strategy if there are at least two agents in the population, a "source" and a "target".

```
sim1 = Simulation(sentencepool="p:20")
# Create p0, p1, ..., p19 in a new local namespace.
# Access them via:
sim1.sentencepool[0], sim1.sentencepool[1],
sim1.sentencepool[-1]
```

A more realistic example There are many settings to the simulation classes, listed below, but not all of them need to be configured for every simulation. Here is an example of a simulation that runs largely on the defaults:

```
# Create 10 positions with strategy attack.
# These will receive random beliefs as none are specified.
my_population = [Position(
    debate=None,
    introduction_strategy=strategies.attack
    ) for _ in range(10)]

# Set up a simulation with a sentence pool of 20, assign
# the population to it and set an argument length to 2
# or 3 premises.
sim2 = Simulation(
    positions=my_population,
```

```
    sentencepool="p:20",
    argumentlength=[2,3]
)
```

B.5.2.2 Simulation types

Iterative argument introductions

class taupy.simulation.simulation.Simulation(directed=True, debate_growth='random', events={'introduction': 9, 'new_sentence': 1}, sentencepool='p:10', max_sentencepool=None, key_statements=None, parent_debate=None, argumentlength=2, positions=None, copy_input_positions=True, initial_position_size=None, default_introduction_strategy={'name': 'random', 'pick_premises_from': None, 'source': False, 'source_accepts_conclusion': 'NA', 'target': False, 'target_accepts_conclusion': 'NA'}, default_update_strategy='closest_coherent', partial_neighbour_search_radius=50)

A simulation in which agents introduce new arguments bit by bit. For historic reasons, this kind of simulation bears the generic name.

Parameters:

`directed (bool)` Boolean indicating whether purposeful argument introductions are requested. If set to False, random arguments that do not take into account agents' belief systems are introduced.

`debate_growth` Can be either of "random" or "treelike". Random debate growth leads to argument maps that are like random graphs. Tree-like debate growth leads to maps that are hierarchical, tree-like structures. The latter require specification of `key_statements`, as the roots of the tree need to be specified.

`events (dict)` A mapping of events to their chance of occuring. Recognised events are "introduction" and "new_sentence".

`sentencepool` The initial pool of sentence available for argument introductions. The input needs to be valid input for `sympy.symbols()`. It is recommded to use sympy's `"p:n"` notation, where `n` refers to the number of sentences.

`max_sentencepool` When the probability of a "new_sentence" event is non-zero, the sentencepool is enlarged in the course of a simulation run until it reaches its maximum extension. As in sentencepool, the input needs to be understood by `sympy.symbols()` and should be equal or greater than sentencepool. If `None` is input, it will default to the extension of sentencepool.

`key_statements` When a "treelike" `debate_growth` is selected, should be an iterable of inputs understood by `sympy.symbols()`. Has no effect when "random" `debate_growth` is selected. These key statements will be the roots of constructed tree-like argument map. Needs to be a subset of elements from sentencepool and `max_sentencepool`.

Example

A sentence pool of `"p:20"` includes the symbols `p0`, `p1`, ..., `p19`. The first two items can be selected as roots for the argument map like this:

```
s = Simulation(debate_growth="treelike",
               sentencepool="p:20",
               key_statements=["p0", "p1"])
```

`parent_debate` If supplied, the simulation will inherit this debate stage. Otherwise, simulations are initialised with an empty debate stage.

`argumentlength` Either an integer or an iterable of integers indicating the number of premises per argument. If an integer *n* is provided, all arguments will have *n* premises. If an iterable is provided, one of its elements *e* is chosen randomly at each argument introduction, and the introduced arguments will receive *e* premises.

Example

Arguments with exactly four premises are introduced to a simulation:

```
s = Simulation(argumentlength=4)
```

Arguments with 1–4 premises are introduced to this simulation:

```
s = Simulation(argumentlength=[1,2,3,4])
```

`positions` An iterable of agents participating in the debate. If not supplied, only random arguments can be introduced in a simulation. Other argumentation strategies require at least two agents in the population.

Example

First generate a list of 10 agents with the fortify strategy:

```
mypositions = [Position(debate=None,
    introduction_strategy=strategies.fortify)
    for _ in range(10)]
```

And then use it in the Simulation initialisation:

```
s = Simulation(positions=mypositions)
```

`copy_input_positions` Decide whether to make a deep copy of the input positions. If set to False, the input position objects will be mutated by the simulation run.

`initial_position_size` If given as an integer *i*, positions will be filled up with random truth-value attributions until they contain *i* such judgements. When None is selected, agents will have complete positions, i.e. they will assign a truth-value to each sentence in the sentence pool.

`default_introduction_strategy` The introduction strategy for positions that have no `introduction_strategy` assigned to them.

`default_update_strategy` Specifies how agents should update their belief system in case of incoherence. Two methods are implemented:

- `"closest_coherent"`: recommended for complete positions
- `"closest_closed_partial_coherent"`: recommended for partial positions

`partial_neighbour_search_radius` A parameter for the `"closest_closed_partial_coherent"`. As the number of partial neighbours rises exponentially when the distance increases, this puts an upper limit on the number of inspected possible positions that an agent with a partial position would move to.

Pre-compiled argument maps

```
class taupy.simulation.simulation.FixedDebateSimulation(argu-
    ment_selection_strategy = 'any', debate_generation =
    {'max_density': 1.0}, default_update_strategy='closest_-
    coherent', initial_arguments=None, initial_position_size
    = None, num_key_statements=1, partial_neighbour_search_-
    radius=100, positions=None, sentencepool='p:10')
```

A simulation that begins with a pre-defined debate. Agents uncover arguments from the debate in each simulation step. The pre-defined debate follows the argument map generation algorithm Betz, Chekan & Mchedlidze (2021).

Paramters:

`argument_selection_strategy` Can be `"any"` or `"max"`. When set to `"any"`, a random argument is uncovered at each step as long as it matches the argumentation strategies' requirements. If `"max"` is selected, arguments that reach the most agents in the population is selected.

`debate_generation (dict)` Additional settings to the initial argument map generation which are passed on to `generate_hierarchical_argument_map()`. Note that the number of key statements and the sentencepool are provided in dedicated attributes `num_key_statements` and `sentencepool`.

`default_update_strategy` Agents who need to update their position do so according to the strategy selected here. Options are `"closest_coherent"` for complete positions and `"closest_closed_partial_coherent"` for partial positions.

`initial_arguments` Start the simulation with these arguments, even if they are not part of the generated argument map. It is advised to use the default None.

`initial_position_size` The number of belief for the initial positions. If set to None, will default to the length of the sentencepool. If smaller than the sentencepool, positions will be partial and contain a random subset of sentences. Positions that have no initial truth-value attributions will be randomly assigned.

`num_key_statements (int)` The number of roots for the generated tree-like argument map.

`partial_neighbour_search_radius (int)` An additional setting for the `"closest_closed_partial_coherent"` updating strategy, indicating the neighbourhood radius that is scanned for an alternative if a position needs to update.

`positions` A list of initial positions, as indicated in Setting up a population.

`sentencepool` A sentencepool, given as an iterable understood by `sympy.symbols()`, to be forwarded to the tree-like argument map generation.

B.5.3 Running a simulation

taupy simulations proceed through introductions or uncoverings of new arguments, agents' reactions to them, and sometimes also introductions of new items to the sentence pool.

The simulation objects `taupy.Simulation` and `taupy.FixedDebateSimulation` have `run()` methods which automatically trigger these events. The `run()` methods accept the following simulation termination conditions:

`max_density` The maximum inferential density, determined by `taupy.Base.density()`, after which a simulation is terminated.

`max_steps` Maximum number of argument introductions, uncoverings, and sentence pool expansion events before the simulation is terminated. Can be set to a high value or to `float("inf")` so that only `max_density` has effect.

`min_sccp` The minimum extension of the set of coherent and complete positions for the debate. If there are fewer positions in the SCCP at a debate stage, the simulation is terminated.

`Simulation.run(max_density=0.8, max_steps=1000, min_sccp=1, quiet=True)`
Run a Simulation using `introduction_method` and `update_mechanism` until either `max_density` is reached, the SCCP has an extension of `min_sccp` or `max_steps` have been taken.

If `quiet=False`, the last log entry which contains a summary of the simulation is not output. This is useful in batch processing of Simulations (see `experiment()`).

`FixedDebateSimulation.run(max_density = 0.8, max_steps = 200, min_sccp = 1, quiet = True)`
Run Simulation steps until targets are reached.

Example

Let's run a simulation `s` until either a density of 0.8 is reached or 200 argument and sentence introductions have been executed, whichever occurs first:

```
s.run(max_density=0.8, max_steps=200)
```

B.5.4 Experiments: Simulations in parallel

As simulations involve random processes it is often necessary to inspect multiple simulation runs to ensure one didn't end up with an outlier. The `experiment()` function can process any number of simulations in parallel.

> **Warning**
>
> Calling `experiment()` will start the requested simulations immediately. By default, it will use all available CPUs on your system to do so. Depending on the set-up, this can keep your machine busy for quite some time. You can employ less CPUs by changing the `max_workers` setting in the executor argument.

taupy.simulation.simulation.experiment(n, *, sim_type=<class 'taupy.simulation.simulation.Simulation'>, executor={}, simulations={}, runs={})

Generate and execute `n` number of Simulations and output their results. The Simulations can be controlled via a dictionary passed to `simulations`. The `Simulation.run()` can be controlled with a dictionary passed to runs.

Settings to the `ProcessPoolExecutor` should be forwarded in a dictionary to executor.

This function calls two Executors. The first is responsible for setting up the Simulations in parallel. The second performs the simulation runs. This is particularly helpful for Simulation types that involve substantial computation for set-up, such as FixedDebateSimulation.

```
# First, create 10 positions with strategy random
positions = [Position(debate=None,
    introduction_strategy=strategies.random)
    for _ in range(10)]

# Run 4 simulations in an experiment (multi-threaded!):
my_experiments = experiment(
    n=4,
    simulations={
        "positions": positions,
        "sentencepool": "p:10",
        "argumentlength": [2,3]
        },
    runs={
        "max_density": 0.8,
        "max_steps": 200
        }
    )
```

This creates an object `my_experiments` with four elements: `my_experiments[0]` contains the first simulation, the second is in `my_experiments[1]`, etc.

The dictionary in the `simulations` argument contains the arguments for creation of the Simulation objects. For example, the above call to `taupy.experiment()` creates simulation object s that look like this:

```
s = Simulation(positions=positions,
               sentencepool="p:10",
               "argumentlength"=[2,3]
               )
```

And the directives in the dictionary `runs` are arguments to the method `Simulation.run()` which is called by the experiments. The settings above are equivalent to:

```
s.run(max_density=0.8, max_steps=200)
```

B.5.5 Evaluation

The `taupy.Evaluation` class provides methods for analysing large chunks of data and storing that information in a combined table. It performs most of its operations concurrently and gives a performance advantage on machines with many CPUs. All measures described in Section B.3 can be applied to `Evaluation` objects.

B.5.5.1 Setting up an Evaluation object

class taupy.simulation.evaluation.Evaluation(*, debate_stages, list_of_positions=None, clustering_method=None, multiprocessing_settings={})

A class to collect measurement values for a simulation while storing shared information between evaluation functions (such as clusterings).

Parameters:

`debate_stages` An iterator containing the lists of debate stages for each simulation run.

`list_of_positions` An iterator containing the lists of belief systems for each simulation run.

`clustering_method` When evaluation functions that rely on position clustering are called, the clustering algorithm specified here will be used. Functions from taupy.analysis.clustering can be selected here, in particular `leiden()`, `affinity_propagation()`, and `agglomerative_clustering()`.

`multiprocessing_settings (dict)` Settings forwarded to multiprocessing. Should be options that are recognised by `concurrent.futures.ProcessPoolExecutor`.

Variables:

`data` A `pandas.DataFrame` containing the analysed data.

B.5.5.2 Viewing results

All measurement functions from the evaluation module are configured to add columns to a shared `pandas.DataFrame` stored in `Evaluation.data`.

```
e = Evaluation()
# View the DataFrame
e.data
# Since e.data is a pandas DataFrame, all DataFrame
# operations can be used:
e.data.to_csv("myexport.csv")
```

An `Evaluation.data` table is structured like in Table B.1.

Table B.1: The first four rows of a hypothetical `Evaluation.data` table that contains data for inferential density and dispersion

		density	dispersion
0	0	0.02324	0.29561402
0	1	0.07451	0.30156791
0	2	0.08462	0.30196067
0	3	0.09880	0.30971113

The first two columns indicate the `pandas.MultiIndex` for the table. The first column corresponds to the simulation number within the experiment, and the second column to the debate stage within the simulation. The remaining columns are inserted by the `Evaluation` class methods described below.

B.5.5.3 A minimal example

Suppose you have run an experiment with iterative argument introductions and want to analyse the density and pairwise dispersion of each debate stage.

```
# First, create 10 positions with strategy random
my_population = [Position(debate=None,
    introduction_strategy=strategies.random)
    for _ in range(10)]

# Run 4 simulations in an experiment:
my_experiments = experiment(
    n=4,
    simulations={
        "positions": my_population,
```

```
        "sentencepool": "p:10",
        "argumentlength": [2,3]
        },
    runs={
        "max_density": 0.8,
        "max_steps": 200
        }
)

# Create an Evaluation object
e = Evaluation(
    debate_stages=my_experiments,
    list_of_positions=[e.positions for e in my_experiments]
)
# Add a density column to the data
e.densities()
# Add a column with dispersion measurements to the data
e.dispersions()
```

The resulting `e.data` table is intended for further data analysis, such as statistics or plotting. These operations will be performed outside of taupy, in Python packages such as `numpy` or `seaborn`.

B.5.5.4 Adding data to an Evaluation object

Shortcut functions These functions are shortcuts to the functions explained in more detail below.

`Evaluation.densities()`
: A shortcut function to directly add the densities to the evaluation DataFrame.

`Evaluation.dispersions(*, configuration={})`
: A shortcut function to directly add pairwise dispersion measurements to the evaluation DataFrame.

`Evaluation.agreement_means(*, configuration={})`
: A shortcut to directly add the mean population-wide agreement to the evaluation DataFrame.

Measures that only analyse debate stages

`Evaluation.debate_stage_analysis(function)`
: A generic evaluation method to analyse, in multiprocessing, only debate stages without taking further data into account. From this module, functions that can be passed to `function` are:

- densities_of_debate_stages()
- sccp_extension()
- progress()

Measures that only analyse positions

Evaluation.position_analysis(*, function, configuration={})
A generic method to evaluate functions that work on positions, with multiprocessing. Examples are (see the shortcut functions as well):

- dispersions_between_positions()
- mean_agreement_between_positions()

taupy.simulation.evaluation.dispersions_between_positions(positions, *, measure=<function normalised_hamming_distance>)

taupy.simulation.evaluation.mean_agreement_between_positions(positions, *, measure=<function bna>)

Measures that rely on clustering

Evaluation.generate_clusters(*, clustering_settings={})
Apply the clustering algorithm selected in Evaluation.clustering_method to the stored debate stages and positions. The clusters are saved in the Evaluation.clusters list and can be accessed by functions that work on clusterings.

Evaluation.group_divergence(*, measure=<function normalised_hamming_distance>)
Calculate the group divergence between all positions stored in the Evaluation object and add a column to the data object. Raises an error if no clustering has been generated.

See taupy.analysis.polarisation.group_divergence() for details.

Evaluation.group_consensus(*, measure=<function normalised_hamming_distance>)
Calculate the group consensus between all positions stored in the Evaluation object and add a column to the data object. Raises an error if no clustering has been generated.

See taupy.analysis.polarisation.group_consensus() for details.

Evaluation.clusters_analysis(*, function, column_name='NAME', configuration={})

Generic multi-process function to apply a measure that works on the cluster structure of a simulation.

Parameters:

`function` A function to be applied in multiprocessing. Here is a list of examples from different taupy submodules that work with this function:

- `number_of_groups`
- `group_size_parity`
- `coverage_of_clustering`
- `Shannon_index`
- `normalised_Shannon_index`
- `Simpson_index`
- `inverse_Simpson_index`
- `Gini_Simpson_index`

Note that `group_divergence()` and `group_consensus()` are calculated with dedicated methods. This is because both functions rely on additional information not present in the clustering alone.

`column_name (str)` Title of the column that is added to the Evaluations data table. Should be indicative of the measure that was applied.

B.6 Examples

B.6.1 How to plot the SCCP

In this tutorial, we will be using the drawing capabilities from `networkx` to plot a simple debate's space of coherent and complete positions (SCCP).

Let's go ahead with a really simple debate:

```
from taupy import *
from sympy import symbols

# declare five sentence variables
p1, p2, p3, p4, p5 = symbols("p:5")

# create a simple debate with two arguments.
d = Debate(Argument(p1&p2&p3, p4), Argument(p5&~p2, ~p3))
```

`taupy.basic.core.Base.sccp()` returns a dictionary of list representation of a graph, which can be easily imported by `networkx`.

```
import networkx as nx

# create the graph with Debate.sccp()
graph = nx.from_dict_of_lists(d.sccp())

# choose a layout (optional)
layout = nx.kamada_kawai_layout(graph)

# plot the graph. networkx uses matplotlib here.
nx.draw(graph, pos=layout, with_labels=True,
        node_color="#aaa")
```

In the resulting Figure B.3, nodes have a label that shows the bit string representation of the position they resemble. In this bit string, propositions are ordered alphabetically (`p1` will show in the first bit, `p5` in the last), and they are either 0 if the position assigns False to this proposition, or 1 if the position accepts the proposition. Positions are connected by an edge if they differ in exactly one truth-value attribution (i.e., if their Hamming distance equals 1).

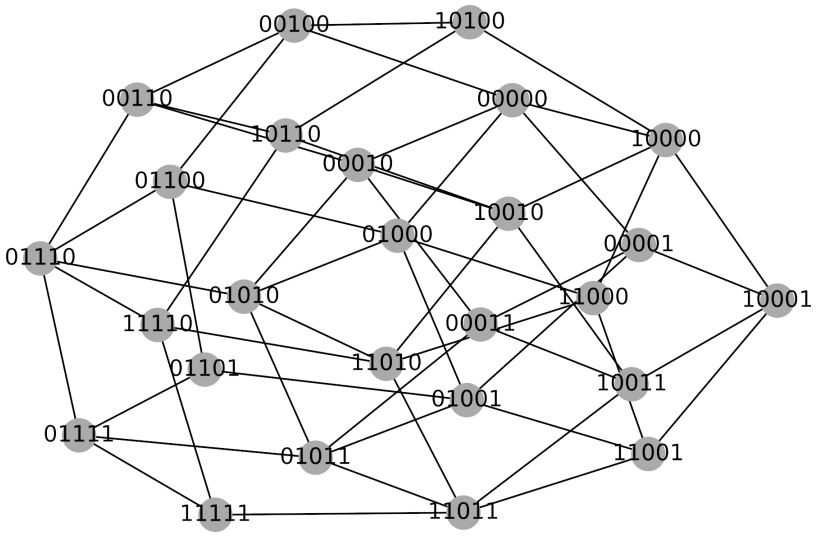


Figure B.3: SCCP plotted for a debate with two arguments, $(p_1 \wedge p_2 \wedge p_3) \Rightarrow p4$ and $(p_5 \wedge \neg p_2) \Rightarrow \neg p_3$).

B.6.2 Track agreement in debate simulations

We want to know whether a pair of agents with randomly initialised belief systems can achieve consensus through argumentation. This experiment reproduces results from Betz (2013, Chapter 6).

Since we also want to understand how different argumentation strategy affect consensus conduciveness, we set up two sets of agents, one equipped with the "attack" and the other with the "convert" strategy.

```
from taupy import *
attack_population = \
    [Position(debate=None,
              introduction_strategy=strategies.attack) \
              for _ in range(2)
    ]

convert_population = \
    [Position(debate=None,
              introduction_strategy=strategies.convert) \
              for _ in range(2)
    ]
```

Next, conduct the experiment with fairly standard simulation settings:

Warning

The following Python code executes 12 simulations in multiprocessing twice. Unless otherwise specified, multiprocessing will use all CPUs available on your machine. On modern consumer hardware, a simulation with 20 sentences takes about 10–30 minutes, depending on the number of agents. If your machine can execute four jobs at the same time, the worst scenario is 180 minutes execution time for both experiments. A machine with 12 CPUs might do the same in less than 30 minutes.

```
attack_experiment = experiment(
    n=12,
    simulations={"positions": attack_population,
                 "sentencepool": "p:20"},
    runs={"max_density": 0.8}
)

convert_experiment = experiment(
    n=12,
    simulations={"positions": convert_population,
                 "sentencepool": "p:20"},
    runs={"max_density": 0.8}
)
```

Let's now analyse all our data. We are interested in the mean population-wide agreement at each debate stage, and the inferential density of that debate

stage – this should tell us whether the continued exchange of arguments leads increases mean agreement between agents, and is thus consensus-conducive.

```
attack_eval = Evaluation(
    debate_stages = attack_experiment,
    list_of_positions = \
        [e.positions for e in attack_experiment]
)

attack_eval.densities()
attack_eval.agreement_means()

convert_eval = Evaluation(
    debate_stages = convert_experiment,
    list_of_positions = \
        [e.positions for e in convert_experiment]
)

convert_eval.densities()
convert_eval.agreement_means()
```

Our raw data is generated! Now let's combine this data and plot it (see Figure B.4 for the result):

```
import pandas as pd
import seaborn

# Add a column to our data indicating which strategy was
# in use:
attack_eval.data["strategy"] = "attack"
convert_eval.data["strategy"] = "convert"

our_data = pd.concat([attack_eval.data, convert_eval.data])
# Convert data to types recognised by seaborn
our_data["agreement"] = \
    our_data["agreement"].astype("float64")
our_data["density"] = \
    our_data["density"].astype("float64").round(2)

# Plot
seaborn.lineplot(
    data=our_data,
    x="density",
    y="agreement",
    hue="strategy"
    )
```

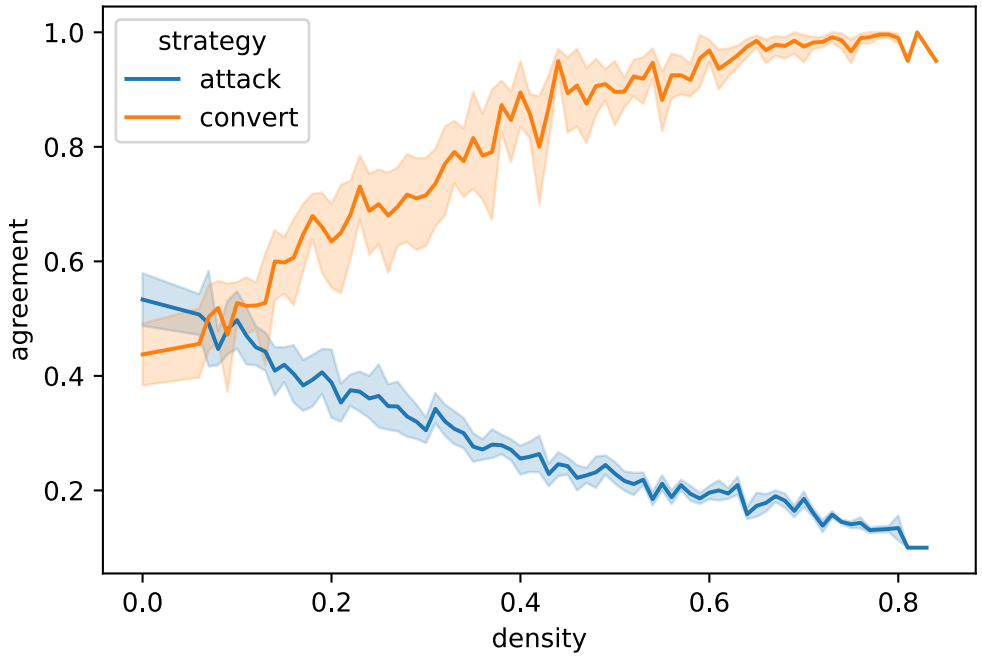


Figure B.4: Mean agreement dynamics in many runs for two different argumentation strategies. Reproduces results from Betz (2013, Chapter 6).

References

Arieli, Ofer, AnneMarie Borg, Jesse Heyninck & Christian Straßer (2021). Logic-based approaches to formal argumentation. In: *Journal of Applied Logics* 8.6, pp. 1793–1898. URL: http://collegepublications.co.uk/ifcolog/?00048.

Banisch, Sven & Eckehard Olbrich (2021). An argument communication model of polarization and ideological alignment. In: *Journal of Artificial Societies and Social Simulation* 24.1. DOI: 10.18564/jasss.4434.

Barnes, Jonathan (1969). Aristotle's theory of demonstration. In: *Phronesis* 14.2, pp. 123–152. URL: https://www.jstor.org/stable/4181832.

Bednar, Jenna (2021). Polarization, diversity, and democratic robustness. In: *Proceedings of the National Academy of Sciences* 118.50, e2113843118. DOI: 10.1073/pnas.2113843118.

Besnard, Philippe & Anthony Hunter (2008). Elements of argumentation. Cambridge, MA: The MIT Press.

Betz, Gregor (2005). The vicious circle theorem: A graph-theoretical analysis of dialectical structures. In: *Argumentation* 19.1, pp. 53–64. DOI: 10.1007/s10503-004-2068-9.

Betz, Gregor (2009). Evaluating dialectical structures. In: *Journal of Philosophical Logic* 38, pp. 283–312. DOI: 10/cxrbhh.

Betz, Gregor (2012). On degrees of justification. In: *Erkenntnis* 77, pp. 237–272. DOI: 10.1007/s10670-011-9314-y.

Betz, Gregor (2013). Debate dynamics: How controversy improves our beliefs. Berlin: Springer. DOI: 10/d3cx.

Betz, Gregor (2022). Natural-language multi-agent simulations of argumentative opinion dynamics. In: *Journal of Artificial Societies and Social Simulation* 25.1. DOI: 10.18564/jasss.4725.

Betz, Gregor, Vera Chekan & Tamara Mchedlidze (2021). *Heuristic algorithms for the approximation of Mutual Coherence*. DOI: 10.48550/arXiv.2307.01639.

Black, Max (1966). The raison d'être of inductive argument. In: *The British Journal for the Philosophy of Science* 17.3, pp. 177–204. DOI: 10.1093/bjps/17.3.177.

Borg, AnneMarie, Daniel Frey, Dunja Šešelja & Christian Straßer (2018). Epistemic effects of scientific interaction: Approaching the question with an argumentative agent-based model. In: *Historical Social Research* 43.1, pp. 285–309. DOI: 10.12759/hsr.43.2018.1.285-309.

Bovens, Luc & Stephan Hartmann (2003). Bayesian epistemology. Oxford: Oxford University Press. DOI: 10.1093/0199269750.001.0001.

Boxell, Levi, Matthew Gentzkow & Jesse M. Shapiro (2024). Cross-country trends in affective polarization. In: *The Review of Economics and Statistics* 106.2, pp. 557–565. DOI: 10.1162/rest_a_01160.

Bramson, Aaron, Patrick Grim, Daniel J. Singer, William J. Berger, Graham Sack, Steven Fisher, Carissa Flocken & Bennett Holman (2017). Understanding polarization: Meanings, measures, and model evaluation. In: *Philosophy of Science* 84.1, pp. 115–159. DOI: 10.1086/688938.

Bramson, Aaron, Patrick Grim, Daniel J. Singer, Steven Fisher, William Berger, Graham Sack & Carissa Flocken (2016). Disambiguation of social polarization concepts and measures. In: *The Journal of Mathematical Sociology* 40.2, pp. 80–111. DOI: 10/d3kn.

Broome, John (2021). Reasons and rationality. In: *The handbook of rationality*. Ed. by Markus Knauff & Wolfgang Spohn. Cambridge, MA: The MIT Press, pp. 129–136. DOI: 10.7551/mitpress/11252.003.0012.

Bruner, Justin P. (2015). Diversity, tolerance, and the social contract. In: *Politics, Philosophy and Economics* 14.4, pp. 429–448. DOI: 10.1177/1470594x14560763.

Burnstein, Eugene & Amiram Vinokur (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. In: *Journal of Experimental Social Psychology* 13.4, pp. 315–332. DOI: 10.1016/0022-1031(77)90002-6.

Butler, George, Gabriella Pigozzi & Juliette Rouchier (2019). Mixing dyadic and deliberative opinion dynamics in an agent-based model of group decision-making. In: *Complexity* Article 3758159. DOI: 10.1155/2019/3758159.

Cartwright, Nancy (2013). Evidence, argument and prediction. In: *EPSA11: Perspectives and foundational problems in philosophy of science*. Ed. by Vassilios Karakostas & Dennis Dieks. The European Philosophy of Science Association Proceedings 2. Cham: Springer, pp. 3–17. DOI: 10.1007/978-3-319-01306-0_1.

Cayrol, Claudette & Marie-Christine Lagasquie-Schiex (2005). On the acceptability of arguments in bipolar argumentation frameworks. In: *Symbolic and quantitative approaches to reasoning with uncertainty, ECSQARU 2005*. Ed. by Lluís Godo. Lecture Notes in Computer Science 3571. Berlin: Springer. DOI: 10/c4qksv.

Comesaña, Juan (2020). Being rational and being right. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198847717.001.0001.

Csárdi, Gábor & Tamás Nepusz (2006). The igraph software package for complex network research. In: *InterJournal Complex Systems* 1695.

de Condorcet, Nicolas (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris: Imprimerie Royale.

Dietrich, Franz (2021). Fully Bayesian aggregation. In: *Journal of Economic Theory* 194. DOI: 10.1016/j.jet.2021.105255.

Dorst, Kevin (2023). Rational polarization. In: *The Philosophical Review* 132.3, pp. 355–458. DOI: 10.1215/00318108-10469499.

Dung, Phan Minh (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. In: *Artificial Intelligence* 77.2, pp. 321–357. DOI: 10/csfr54.

Dutilh Novaes, Catarina (2021). Argument and argumentation. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2021/entries/argument/.

Dutilh Novaes, Catarina (2023). Can arguments change minds? In: *Proceedings of the Aristotelian Society* 123.2, pp. 173–198. DOI: 10.1093/arisoc/aoad006.

Elsenbroich, Corinna (2012). Explanation in agent-based modelling: Functions, causality or mechanisms? In: *Journal of Artificial Societies and Social Simulation* 15.3. DOI: 10.18564/jasss.1958.

Esteban, Joan-María & Debraj Ray (1994). On the measurement of polarization. In: *Econometria* 62.4, pp. 819–851.

Feldman, Richard (2014). Reason and argument. 2nd ed. Harlow: Pearson Education.

Flache, Andreas, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet & Jan Lorenz (2017). Models of social influence: Towards the next frontiers. In: *Journal of Artificial Societies and Social Simulation* 20.4. DOI: 10.18564/jasss.3521.

Fogal, Daniel & Alex Worsnip (2021). Which reason? Which rationality? In: *Ergo* 8. DOI: 10.3998/ergo.1148.

Frey, Brendan J. & Delbert Dueck (2007). Clustering by passing messages between data points. In: *Science* 315.5814, pp. 972–976. DOI: 10.1126/science.1136800.

Frey, Daniel & Dunja Šešelja (2018). What is the epistemic function of highly idealized agent-based models of scientific inquiry? In: *Philosophy of the Social Sciences* 48.4: *Selected papers from the 6th ENPOSS meeting, Cracow, 20–22 September, 2017*. Ed. by Tomasz Kwarcinski et al., pp. 407–433. DOI: 10.1177/0048393118767085.

Friedkin, Noah E., Anton V. Proskurnikov, Roberto Tempo & Sergey E. Parsegov (2016). Network science on belief system dynamics under logic constraints. In: *Science* 354.6310, pp. 321–326. DOI: 10.1126/science.aag2624.

Friedman, Jane (2013). Suspended judgment. In: *Philosophical Studies* 162, pp. 162–181. DOI: 10.1007/s11098-011-9753-y.

Friedman, Jane (2020). The epistemic and the zetetic. In: *The Philosophical Review* 129.4, pp. 501–536. DOI: 10.1215/00318108-8540918.

Gabbriellini, Simone & Paolo Torroni (2014). A new framework for ABMs based on argumentative reasoning. In: *Advances in social simulation*. Ed. by Bogumił Kamiński & Grzegorz Koloch. Advances in Intelligent Systems and Computing 229. Berlin: Springer, pp. 25–36. DOI: 10.1007/978-3-642-39829-2_3.

Gärdenfors, Peter (1992). Belief revision: An introduction. In: *Belief revision*. Ed. by Peter Gärdenfors. Cambridge Tracts in Theoretical Computer Science 29. Cambridge, UK: Cambridge University Press, pp. 1–28.

Graves, John C. (1974). Uniformity and induction. In: *The British Journal for the Philosophy of Science* 25.4, pp. 301–318. DOI: 10.1093/bjps/25.4.301.

Grim, Patrick, Frank Seidl, Calum McNamara, Hinton E. Rago, Isabell N. Astor, Caroline Diaso & Peter Ryner (2022). Scientific theories as Bayesian nets: Structure and evidence sensitivity. In: *Philosophy of Science* 89.1, pp. 42–69. DOI: 10.1017/psa.2021.18.

Grim, Patrick & Daniel Singer (2024). Computational philosophy. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta & Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2024/entries/computational-philosophy/.

Grim, Patrick, Daniel J. Singer, Aaron Bramson, Bennett Holman, Sean McGeehan & William J. Berger (2019). Diversity, ability, and expertise in epistemic communities. In: *Philosophy of Science* 86, pp. 98–123. DOI: 10.1086/701070.

Grimm, Volker, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K. Heinz, Geir Huse, Andreas Huth, Jane U. Jepsen, Christian Jørgensen, Wolf M. Mooij, Birgit Müller, Guy Pe'er, Cyril Piou, Steven F. Railsback, Andrew M. Robbins, Martha M. Robbins, Eva Rossmanith, Nadja Rüger, Espen Strand, Sami Souissi, Richard A. Stillman, Rune Vabø, Ute Visser & Donald L. DeAngelis (2006). A standard protocol for describing individual-based and agent-based models. In: *Ecological Modelling* 198.1, pp. 115–126. DOI: 10.1016/j.ecolmodel.2006.04.023.

Grüne-Yanoff, Till (2009). The explanatory potential of artificial societies. In: *Synthese* 169, pp. 539–555. DOI: 10.1007/s11229-008-9429-0.

Hallam, Anthony (1989). Great geological controversies. 2nd edition. Oxford: Oxford University Press.

Hartmann, Stephan, Carlo Martini & Jan Sprenger (2009). Consensual decision-making among epistemic peers. In: *Episteme* 6.2, pp. 110–129. DOI: 10.3366/E1742360009000598.

Hegselmann, Rainer & Ulrich Krause (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. In: *Journal of Artificial Societies and Social Simulation* 5.3. URL: https://www.jasss.org/5/3/2.html.

Hegselmann, Rainer & Ulrich Krause (2009). Deliberative exchange, truth, and cognitive division of labour: A low-resolution modeling approach. In: *Episteme* 6.2, pp. 130–144. DOI: 10.3366/E1742360009000604.

Heinzelmann, Nora (2022). Rationality is not coherence. In: *The Philosophical Quarterly* 74.1, pp. 312–332. DOI: 10.1093/pq/pqac083.

Hintikka, Jaakko (1988). Advice to prospective philosophers. In: *Proceedings and Addresses of the American Philosophical Association* 62.1, Supplement, pp. 272–273. URL: https://www.jstor.org/stable/44079724.

Hong, Lu & Scott E. Page (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. In: *Proceedings of the National Academy of Sciences* 101.46, pp. 16385–16389. DOI: 10.1073/pnas.0403723101.

Hubert, Lawrence & Phipps Arabie (1985). Comparing partitions. In: *Journal of Classification* 2, pp. 193–218. DOI: 10.1007/BF01908075.

Hume, David (1741). Of parties in general. In: *Essays, moral and political*. Vol. 1. Edinburgh: A. Kincaid, pp. 105–117.

Hunter, Anthony, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux & Sylwia Polberg (2019). Towards computational persuasion via natural language argumentation dialogues. In: *KI 2019: Advances in Artificial Intelligence*. Ed. by Christoph Benzmüller & Heiner Stuckenschmidt. Lecture Notes in Computer Science 11793. Cham: Springer, pp. 18–33. DOI: 10/gwft.

Indraccolo, Lisa (2021). Argumentation and persuasion in Classical Chinese literature. In: *Essays on argumentation in antiquity*. Ed. by Joseph A. Bjelde, David Merry & Christopher Roser. Argumentation Library 39. Cham: Springer, pp. 21–48. DOI: 10.1007/978-3-030-70817-7_2.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra & Sean J. Westwood (2019). The origins and consequences of affective polarization in the United States. In: *Annual Review of Political Science* 22, pp. 129–146. DOI: 10.1146/annurev-polisci-051117-073034.

Iyengar, Shanto, Gaurav Sood & Yphtach Lelkes (2012). Affect, not ideology: A social identity perspective on polarization. In: *Public Opinion Quarterly* 76.3 (Fall 2012), pp. 405–431. DOI: 10.1093/poq/nfs038.

Iyengar, Shanto & Sean J. Westwood (2015). Fear and loathing across party lines: New evidence on group polarization. In: *American Journal of Political Science* 59.3, pp. 690–707. DOI: 10.1111/ajps.12152.

Jonides, John, Richard L. Lewis, Derek Evan Nee, Cindy A. Lustig, Marc G. Berman & Katherine Sledge Moore (2008). The mind and brain of short-term memory. In: *Annual Review of Psychology* 59.1, pp. 193–224. DOI: 10.1146/annurev.psych.59.103006.093615.

Kahan, Dan M. (2013). Ideology, motivated reasoning, and cognitive reflection. In: *Judgment and Decision Making* 8.4, pp. 407–424. DOI: 10.2139/ssrn.2182588.

Keijzer, Marijn A., Michael Mäs & Andreas Flache (2018). Communication in online social networks fosters cultural isolation. In: *Complexity* Article 9502872. DOI: 10.1155/2018/9502872.

Kiesewetter, Benjamin & Alex Worsnip (2023). Structural rationality. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta & Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2023/entries/rationality-structural/.

Kitcher, Philip (1990). The division of cognitive labor. In: *The Journal of Philosophy* 87.1, pp. 5–22.

Knauff, Markus & Wolfgang Spohn (2021). Psychological and philosophical frameworks of rationality: A systematic introduction. In: *The handbook of rationality*. Ed. by Markus Knauff & Wolfgang Spohn. Cambridge, MA: The MIT Press, pp. 1–65. DOI: 10.7551/mitpress/11252.003.0004.

Kopecky, Felix (2021–2022). *taupy: A Python package to study the theory of dialectical structures*. DOI: 10.5281/zenodo.5067834.

Kopecky, Felix (2022). Arguments as drivers of issue polarisation in debates among artificial agents. In: *Journal of Artificial Societies and Social Simulation* 25.1. DOI: 10.18564/jasss.4767.

Kopecky, Felix (2024). Argumentation-induced rational issue polarisation. In: *Philosophical Studies* 181.1, pp. 83–107. DOI: 10.1007/s11098-023-02059-6.

Kopecky, Felix & Gregor Betz (2025). Inconsistent belief aggregation in diverse and polarised groups. In: *Philosophy of Science* 92.1, pp. 40–58. DOI: 10.1017/psa.2024.29.

Lehrer, Keith (1976). When rational disagreement is impossible. In: *Noûs* 10, pp. 327–332.

Lehrer, Keith & Carl Wagner (1981). Rational consensus in science and society: A philosophical and mathematical study. Philosophical Studies Series 24. Dordrecht: D. Reidel. DOI: 10.1007/978-94-009-8520-9.

List, Christian (2005). The probability of inconsistencies in complex collective decisions. In: *Social Choice and Welfare* 24, pp. 3–32. DOI: 10.1007/s00355-003-0253-7.

List, Christian & Philip Pettit (2002). Aggregating sets of judgments: An impossibility result. In: *Economics & Philosophy* 18.1, pp. 89–110. DOI: 10.1017/S0266267102001098.

Lord, Errol (2013). The importance of being rational. PhD thesis. Princeton University. URL: https://philpapers.org/rec/LORTIO-3.

Lorenz, Kuno (2008). Features of Indian logic. In: *Dialogues, logics and other strange things: Essays in honour of Shahid Rahman.* Ed. by Cédric Dégremont, Laurent Keiff & Helge Rückert. London: College Publications, pp. 92–106.

Madison, James (1787). The same subject continued: The union as a safeguard against domestic faction and insurrection. In: *New York Packet* 23 November 1787.

Mäs, Michael & Andreas Flache (2013). Differentiation without distancing: Explaining bi-polarization of opinions without negative influence. In: *PLOS ONE* 8.11, e74516. DOI: 10.1371/journal.pone.0074516.

Mason, Lilliana (2013). The rise of uncivil agreement: Issue versus behavioral polarization in the American electorate. In: *American Behavioral Scientist* 57.1, pp. 140–159. DOI: 10.1177/0002764212463363.

Mason, Lilliana (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. In: *American Journal of Political Science* 59.1, pp. 128–145. DOI: 10.1111/ajps.12089.

Mayo-Wilson, Conor & Kevin J. S. Zollman (2021). The computational philosophy: Simulation as a core philosophical method. In: *Synthese* 199, pp. 3647–3673. DOI: 10.1007/s11229-020-02950-3.

Mayo-Wilson, Conor, Kevin J. S. Zollman & David Danks (2011). The independence thesis: When individual and social epistemology diverge. In: *Philosophy of Science* 78.4, pp. 653–677. DOI: 10.1086/661777.

McPherson, Miller, Lynn Smith-Lovin & James M Cook (2001). Birds of a feather: Homophily in social networks. In: *Annual Review of Sociology* 27.1, pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415.

Mercier, Hugo & Dan Sperber (2011). Why do humans reason? Arguments for an argumentative theory. In: *Behavioral and Brain Sciences* 34, pp. 57–111. DOI: 10.1017/S0140525X10000968.

Merdes, Christoph, Momme von Sydow & Ulrike Hahn (2021). Formal models of source reliability. In: *Synthese* 198 (Supplement 23), pp. 5773–5801. DOI: 10.1007/s11229-020-02595-2.

Mill, John S. (1843). A system of logic. Vol. 1. London: John W. Parker.

Mutz, Diana C. (2002). Cross-cutting social networks: Testing democratic theory in practice. In: *American Political Science Review* 96.1, pp. 111–126. DOI: 10.1017/S0003055402004264.

Myers, David G. (1975). Discussion-induced attitude polarization. In: *Human Relations* 28.8, pp. 699–714. DOI: 10.1177/001872677502800802.

Nakamura, Shuji, Stephen Pearton & Gerhard Fasol (2000). The blue laser diode: The complete story. 2nd ed. Berlin: Springer. DOI: 10.1007/978-3-662-04156-7.

Nehring, Klaus & Clemens Puppe (2002). A theory of diversity. In: *Econometrica* 70.3, pp. 1155–1198. DOI: 10.1111/1468-0262.00321.

Nehring, Klaus & Clemens Puppe (2009). Diversity. In: *The handbook of rational and social choice*. Ed. by Paul Anand, Prasanta Pattanaik & Clemens Puppe. Oxford: Oxford University Press, pp. 298–320. DOI: 10.1093/acprof:oso/9780199290420.003.0013.

O'Connor, Cailin & James Owen Weatherall (2018). Scientific polarization. In: *European Journal for Philosophy of Science* 8, pp. 855–875. DOI: 10.1007/s13194-018-0213-9.

Olsson, Erik J. (2013). A Bayesian simulation model of group deliberation and polarization. In: *Bayesian argumentation: The practical side of probability*. Ed. by Frank Zenker. Synthese Library 362. Dordrecht: Springer. DOI: 10/ggz2.

Page, Scott E. (2011). Diversity and complexity. Primers in Complex Systems 2. Princeton University Press. DOI: 10.1515/9781400835140.

Pallavicini, Josefine, Bjørn Hallsson & Klemens Kappel (2021). Polarization in groups of Bayesian agents. In: *Synthese* 198, pp. 1–55. DOI: 10.1007/s11229-018-01978-w.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay (2011). Scikit-learn: Machine learning in Python. In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

Pettit, Philip (2001). Deliberative democracy and the discursive dilemma. In: *Philosophical Issues* 11, pp. 267–299.

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

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

Polberg, Sylwia & Anthony Hunter (2018). Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. In: *International Journal of Approximate Reasoning* 93, pp. 487–543. DOI: 10.1016/j.ijar.2017.11.009.

Popper, Karl (1976). The myth of the framework. In: *Rational changes in science: Essays on scientific reasoning*. Ed. by Joseph C. Pitt & Marcello Pera. Boston Studies in the Philosophy of Science 98. Dordrecht: D. Reidel, pp. 35–62.

Pöyhönen, Samuli (2017). Value of cognitive diversity in science. In: *Synthese* 194, pp. 4519–4540. DOI: 10.1007/s11229-016-1147-4.

Quine, Willard van Orman & Joseph S. Ullian (1978). The web of belief. 2nd ed. McGraw-Hill.

Reiter, Raymond (1980). A logic for default reasoning. In: *Artificial Intelligence* 13.1–2, pp. 81–132. DOI: 10.1016/0004-3702(80)90014-4.

Richardson, Henry S. (2002). Democratic autonomy: Public reasoning about the ends of policy. Oxford: Oxford University Press.

Roose, Jochen (2021). Politische Polarisierung in Deutschland: Repräsentative Studie zu Zusammenhalt in der Gesellschaft [Political polarisation in Germany: Representative study on societal cohesion]. Berlin: Konrad-Adenauer-Stiftung. URL: https://www.kas.de/de/einzeltitel/-/content/politische-polarisierung-in-deutschland.

Rosenstock, Sarita, Justin Bruner & Cailin O'Connor (2017). In epistemic networks, is less really more? In: *Philosophy of Science* 84.2, pp. 234–252. DOI: 10.1086/690717.

Russell, Bertrand (1903). The principles of mathematics. Cambridge, UK: Cambridge University Press.

Schoenfield, Miriam (2014). Permission to believe: Why permissivism is true and what it tells us about irrelevant influences on belief. In: *Noûs* 48.2, pp. 193–218. DOI: 10.1111/nous.12006.

Schuster, Daniela (2022). *Forms and norms of indecision in argumentation theory*. Presented at the 15th international conference on deontic logic and normative systems, DEON 2020/2021. DOI: 10.48550/arXiv.2203.02207.

Šešelja, Dunja (2022). Agent-based models of scientific interaction. In: *Philosophy Compass* 17.7. DOI: 10.1111/phc3.12855.

Šešelja, Dunja (2023). Agent-based modeling in the philosophy of science. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta & Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2023/entries/agent-modeling-philscience/.

Shi, Feng, Misha Teplitskiy, Eamon Duede & James A. Evans (2019). The wisdom of polarized crowds. In: *Nature Human Behaviour* 3, pp. 329–336. DOI: 10/c286.

Singer, Daniel J. (2019). Diversity, not randomness, trumps ability. In: *Philosophy of Science* 86.1, pp. 178–191. DOI: 10.1086/701074.

Singer, Daniel J., Aaron Bramson, Patrick Grim, Bennett Holman, Jiin Jung, Karen Kovaka, Anika Ranginani & William J. Berger (2019). Rational social and political polarization. In: *Philosophical Studies* 176.9, pp. 2243–2267. DOI: 10.1007/s11098-018-1124-5.

Straßer, Christian & G. Aldo Antonelli (2019). Non-monotonic logic. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta & Uri Nodelman.

Summer 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2024/entries/logic-nonmonotonic/.

Striker, Gisela (2022). Aristotle's three theories of argument. In: *From Aristotle to Cicero: Essays in ancient philosophy*. Oxford: Oxford University Press, pp. 77–87. DOI: 10.1093/oso/9780198868385.003.0006.

Sunstein, Cass R. (2002). The law of group polarization. In: *The Journal of Political Philosophy* 10.2, pp. 175–195.

Taber, Charles S. & Milton Lodge (2006). Motivated skepticism in the evaluation of political beliefs. In: *American Journal of Political Science* 50.3, pp. 755–769. DOI: 10.1111/j.1540-5907.2006.00214.x.

Thagard, Paul (2018). Computational models in science and philosophy. In: *Introduction to formal philosophy*. Ed. by Sven Ove Hansson, Vincent F. Hendricks & Esther M. Kjeldahl. Cham: Springer, pp. 457–467. DOI: 10.1007/978-3-319-77434-3_24.

Thompson, Abigail (2014). Does diversity trump ability? An example of the misuse of mathematics in the social sciences. In: *Notices of the American Mathematical Society* 61.9, pp. 1024–1030. DOI: 10.1090/noti1163.

Thorstad, David (2021). Inquiry and the epistemic. In: *Philosophical Studies* 178.9, pp. 2913–2928. DOI: 10.1007/s11098-020-01592-y.

Titelbaum, Michael G. (2022). Fundamentals of Bayesian epistemology. Vol. 1: Introducing credences. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198707608.001.0001.

Traag, Vincent A., Ludo Waltman & Nees J. van Eck (2019). From Louvain to Leiden: Guaranteeing well-connected communities. In: *Scientific Reports* 9, p. 5233. DOI: 10/gfxg2v.

Tuomisto, Hanna (2010). A consistent terminology for quantifying species diversity? Yes, it does exist. In: *Oecologia* 164, pp. 853–860. DOI: 10.1007/s00442-010-1812-0.

Vallier, Kevin (2022). Public justification. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta & Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2022/entries/justification-public/.

Van Hees, Martin (2004). Freedom of choice and diversity of options: Some difficulties. In: *Social Choice and Welfare* 22, pp. 253–266. DOI: 10.1007/s00355-003-0285-z.

Vorobej, Mark (2006). A theory of argument. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511498879.

Walton, Douglas, Chris Reed & Fabrizio Macagno (2008). Argumentation schemes. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511802034.

Weisberg, Michael & Ryan Muldoon (2009). Epistemic landscapes and the division of cognitive labor. In: *Philosophy of Science* 76.2, pp. 225–252. DOI: 10.1086/644786.

Weitzman, Martin L. (1992). On diversity. In: *Quarterly Journal of Economics* 107.2, pp. 363–405. DOI: 10.2307/2118476.

Williamson, Timothy (2017). Model-building in philosophy. In: *Philosophy's future: The problem of philosophical progress*. Ed. by Russell Blackford & Damien Broderick. Malden, MA: Wiley, pp. 159–171. DOI: 10.1002/9781119210115.ch12.

Winsberg, Eric (2022). Computer simulations in science. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta & Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2022/entries/simulations-science/.

Zinke, Alexandra (2021). Rational suspension. In: *Theoria* 87.5, pp. 1050–1066. DOI: 10.1111/theo.12320.

Zollman, Kevin J. S. (2007). The communication structure of epistemic communities. In: *Philosophy of Science* 74.5, pp. 574–587. DOI: 10.1086/525605.

Zollman, Kevin J. S. (2010). The epistemic benefit of transient diversity. In: *Erkenntnis* 72, pp. 17–35. DOI: 10.1007/s10670-009-9194-6.