



Social structures and deliberation : exploring socio-technical controversies using abstract argumentation

Louise Dupuis

► To cite this version:

Louise Dupuis. Social structures and deliberation : exploring socio-technical controversies using abstract argumentation. Multiagent Systems [cs.MA]. Université Paris sciences et lettres, 2024. English. ⟨NNT : 2024UP-SLD057⟩. ⟨tel-05021492⟩

HAL Id: tel-05021492

<https://theses.hal.science/tel-05021492v1>

Submitted on 4 Apr 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Université Paris-Dauphine

Structures sociales et délibération

**Etudier les controverses socio-techniques avec l'argumentation
abstraite**

Soutenue par

Louise DUPUIS DE TARLÉ

Le 9 décembre 2024

Ecole doctorale n° ED 543

Ecole doctorale SDOSE

Spécialité

Informatique

Composition du jury :

Nicolas MAUDET
Professeur des universités,
Sorbonne Université

Président

Pascale ZARATÉ
Professeur des universités,
Université Toulouse 1 Capitole

Rapporteuse

Srdjan VESIC
Chargé de recherche,
CRIL, Université d'Artois

Rapporteur

Elise BONZON
Maîtresse de conférence,
Université Paris-Cité

Examinatrice

Sylvie HUET
Ingénieure de recherche,
INRAE Laboratoire
d'Ingénierie des Systèmes
Complexes

Examinatrice

Rallou THOMOPOULOS
Directrice de recherche,
IATE, Université de
Montpellier, INRAE

Examinatrice

Gabriella PIGOZZI
Maîtresse de Conférence,
Université Paris-Dauphine - PSL

Co-encadrante

Juliette ROUCHIER
Directrice de recherche,
Université Paris-Dauphine - PSL

Directrice de thèse

Abstract

From debates about vaccines to the questioning of global warming, from discussions on pesticides to those on the electric car, our societies are regularly crossed by in-depth debates that mobilize scientific, technical, legal, moral, and political notions. Sociologists and philosophers of science such as Bruno Latour call these debates **socio-technical controversies**, and have shown that studying them is essential to understand the trajectory of our societies. As argumentative discussions, these controversies can be explored through the prism of argumentation. In particular, **abstract argumentation** is a field of computer science which simplifies the complex nature of argumentative interactions by representing arguments as abstract objects. This simplification allows us to use formal representation to gain insights about argumentative discussions. Thus, we can use computational models such as agent-based models (ABMs) to investigate the social dynamics of debates.

The aim of my thesis is to **understand the dynamics of socio-technical controversies using abstract argumentation**. In particular, I'm interested in how **social structures** influence the **circulation of arguments** within these controversies. In my work, I explore this question using several approaches: both theoretical and applied, focusing either on the scientific side or on the socio-political side of the controversies.

This manuscript first presents two agent-based models of scientific communities which use abstract argumentation to represent agents' cognition and the circulation of arguments in social and public spaces. One of them uses a novel gradual semantics which was developed specifically to model agents' opinions. These models are innovative while being part of the tradition of formal social epistemology and, more broadly, of opinion dynamics. They investigate the role of social structures: the shape of the social network, and the presence of filters of weak arguments, on the outcome of scientific debates.

The second part of this thesis presents real-life case studies of socio-technical controversies using abstract argumentation. The first is a debate between proponents and opponents to the construction of a large sawmill in Tarn, France. Using a visualisation tool and interviews, I represented the controversy as an abstract argumentation graph, gathering insights about the argumentative strategies of both camps. The second case study, investigates the scientific arguments put forward by climate contrarians (actors who promote skepticism about climate change), showing their internal contradictions and the misuse of scientific evidence as an argumentative strategy.

Together these works, both simulations and real-case studies, paint a picture of controversies as dynamic processes, subject to manipulation and whose outcome is heavily influenced by social structures. They show the relevance of formal tools, and in particular abstract argumentation, to study these questions.

Keywords: Argumentation · Controversies · Agent-Based Models · Opinion Dynamics · Social Epistemology

Résumé

Des débats sur les vaccins aux interrogations sur le réchauffement climatique, des discussions sur les pesticides à celles sur la voiture électrique, nos sociétés sont régulièrement traversées par des débats de fond qui mobilisent des notions scientifiques, techniques, juridiques, morales et politiques. Les sociologues et philosophes des sciences comme Bruno Latour appellent ces débats des "controverses socio-techniques", dont l'étude est essentielle pour comprendre la trajectoire de nos sociétés. En tant que discussions argumentées, ces controverses peuvent être explorées à travers le prisme de l'argumentation. En particulier, **l'argumentation abstraite** est un domaine de l'informatique qui simplifie la nature complexe des interactions argumentatives en représentant les arguments comme des objets abstraits. Cette simplification nous permet d'utiliser une représentation formelle pour mieux comprendre les discussions argumentatives. Nous pouvons ainsi utiliser des modèles informatiques tels que les modèles à base d'agents (ABM) pour étudier la dynamique sociale des débats.

L'objectif de ma thèse est de comprendre **la dynamique des controverses socio-techniques** en utilisant **l'argumentation abstraite**. En particulier, je m'intéresse à la manière dont les **structures sociales** influencent la **circulation des arguments** au sein de ces controverses. Dans mon travail, j'explore cette question en utilisant plusieurs approches : à la fois théoriques et appliquées, en me concentrant soit sur l'aspect scientifique, soit sur l'aspect sociopolitique des controverses.

Ce manuscrit présente d'abord deux modèles de communautés scientifiques basés sur des agents qui utilisent l'argumentation abstraite pour représenter leur cognition et la circulation des arguments dans les espaces sociaux et publics. L'un d'entre eux utilise une nouvelle sémantique graduelle qui a été développée spécifiquement pour modéliser les opinions des agents. Ces modèles sont innovants tout en s'inscrivant dans la tradition de l'épistémologie sociale formelle et, plus largement, de la dynamique des opinions. Ils étudient le rôle des structures sociales : la forme du réseau social, et la présence de filtres d'arguments faibles, sur l'issue des débats scientifiques.

La deuxième partie de cette thèse présente des études de cas réels de controverses socio-techniques utilisant l'argumentation abstraite. La première est un débat entre les partisans et les opposants à la construction d'une grande scierie dans le Tarn, en France. À l'aide d'un outil de visualisation et d'entretiens, j'ai représenté la controverse sous la forme d'un graphe d'argumentation abstraite, ce qui m'a permis de recueillir des informations sur les stratégies argumentatives des deux camps. La seconde étude de cas porte sur les arguments scientifiques avancés par les opposants au changement climatique en montrant leurs contradictions internes et l'utilisation abusive des preuves scientifiques comme stratégie d'argumentation.

L'ensemble de ces travaux, qu'il s'agisse de simulations ou d'études de cas réels,

dépeint les controverses comme des processus dynamiques, sujets à manipulation et dont l'issue est fortement influencée par les structures sociales. Il montre la pertinence des outils formels, et en particulier de l'argumentation abstraite, pour étudier ces questions.

Mots-clés: Argumentation · Controverses · Modèles multi-agents · Dynamiques des opinions · Epistémologie sociale

Contents

1	Introduction	1
1.1	Formal argumentation	2
1.2	Agent-based models	4
1.3	Social epistemology	5
1.4	Research questions and method	6
1.5	Thesis structure	7
I	Context and Tools	9
2	Abstract Argumentation	11
2.1	Acceptability Semantics	12
2.2	Bipolar Graphs	13
2.3	Gradual Semantics	15
2.4	Empirical Validation of Abstract Argumentation	16
2.4.1	Testing set-based semantics	17
2.4.2	Testing fine-grained approaches	18
2.5	Comparing Abstract Argumentation Approaches	20
2.5.1	Fine grained or set-based semantics	20
2.5.2	Bipolar or attack graph	21
2.5.3	Gradual or epistemic approach	21
2.5.4	Weighted or non weighted graphs	22
2.6	Multi-Agent Protocols	23
2.6.1	Team persuasion	23
2.6.2	Multiparty persuasion	24
2.6.3	Conclusion	26
3	Agent-Based Models	29
3.1	Generalities on Agent-Based Models	29
3.1.1	Building a model	31
3.1.2	Epistemology of ABMs	31
3.1.3	Verification and validation of ABMs	32
3.2	Opinion Dynamics	34
3.2.1	Key concepts	35
3.2.2	Classical models of opinion	37
3.2.3	Modelling opinion with arguments	39
3.3	Models of Epistemic Communities	41
3.3.1	Truth in the bounded confidence model	42
3.3.2	Bayesian networks	44
3.3.3	ArgABM: modelling scientific inquiry with argumentation	49

4	Agent-Based Models of Abstract Argumentation	51
4.1	Knowledge Representation	53
4.1.1	Mental argumentation frameworks	53
4.1.2	Forming positions on a debate	58
4.2	Social Aspects	60
4.3	Results	63
4.4	Conclusion	64
II	Modelling Scientific Controversies with Abstract Ar-	65
	gumentation	
	Introduction to Part II	67
5	A Gradual Semantics to Model Opinion with Bipolar Argumenta-	69
	tion Frameworks	
5.1	Introduction	69
5.2	Opinion Model	70
5.3	Desirable Properties	72
5.3.1	First principles	72
5.3.2	EBS and the opinion model	73
5.3.3	Open Mindedness and Duality	75
5.4	Novel Semantics	77
5.4.1	Convergence of Opinions	79
5.5	Conclusion	81
6	Exploring Opinion Diversity and Epistemic Success in an Argu-	83
	mentative Model	
6.1	Introduction	83
6.2	Abstract Argumentation	84
6.2.1	Bipolar argumentation frameworks	84
6.2.2	Logistic sum semantic for bipolar graphs	85
6.3	The Model	86
6.3.1	Setting	86
6.3.2	Protocol	86
6.3.3	Examples	88
6.4	Results	90
6.4.1	Stability	90
6.4.2	Opinion convergence	92
6.4.3	Epistemic success	95
6.5	Discussion and Conclusion	97
7	An Agent-Based Model of Myside Bias in Scientific Debates	99
7.1	Introduction	99
7.2	Myside Bias and Scientific Argumentation	102
7.2.1	Myside bias as a production bias	102
7.2.2	Abstract argumentation	103
7.3	The Model	105

7.3.1	Modelling opposing views with argumentation frameworks . . .	105
7.3.2	Baseline protocol	108
7.3.3	MySide Bias in the production of arguments	110
7.3.4	Shared beliefs as shared standards of evaluation	112
7.4	Simulations and Observations	113
7.4.1	Epistemic evaluation	114
7.5	Results	116
7.5.1	Time convergence	116
7.5.2	The baseline model	117
7.5.3	A biased community	119
7.5.4	Results with shared beliefs	123
7.6	Related Works	125
7.7	Conclusion and Outlook	126
8	A Brief Synthesis of the Models	129
III	Exploring Applications	133
	Introduction to Part III	135
9	Visualizing Argumentative Strategies in Controversies over Con-	137
	tested Projects	
9.1	Introduction	137
9.2	Context	138
9.2.1	The "mega-sawmill" project	138
9.3	Method	139
9.3.1	A tool to represent argumentation graphs	139
9.3.2	Co-construction of the argumentation framework of the oppo- nents	139
9.3.3	Creating the argumentation framework of the supporters of the project	141
9.4	Results	141
9.4.1	The CON graph	141
9.4.2	The PRO graph	144
9.4.3	Merging the graphs	145
9.4.4	Informing argumentative strategies	145
9.5	Related Works	148
9.5.1	Empirical studies of abstract argumentation	149
9.5.2	The study of controversies	150
9.6	Discussion and Conclusion	152
10	"Casting Doubt on Consensus": The Use of Scientific Arguments	155
	to Promote Climate Contrarianism	
10.1	Introduction	155
10.2	An Argumentation Framework of Climate Skepticism	159
10.3	Qualitative Study of Climate Skeptics Blogs	163
10.3.1	Description	164

10.3.2 Misuse of articles	164
10.4 Discussion and Future Developments	166
11 Conclusion	169
Remerciements - Aknowledgements	179
Résumé de la Thèse en Français	183
Bibliography	201
Appendix	.1
A Appendix of Chapter 5	.1
A.1 Open-Mindedness	.1
A.2 Duality	.2
B Appendix of Chapter 7	.4
B.1 Algorithm for computing the acceptable arguments	.4
B.2 Additional results	.4
B.3 Proof of Observation 1	.5

CHAPTER 1

Introduction

Would you want to drink a bowl of RoundUp? Surprisingly, science has not settled the matter concerning the toxicity of glyphosate, the main active component of the well-known Monsanto herbicide (Caiati et al., 2020). This controversy peaked in 2015 with the classification of the molecule as "probably carcinogenic" by the International Agency for Research on Cancer, with conflicting evidence pointing to both sides, and the reveal of the "Monsanto papers", a collection of internal documents showing the corruption of scientists by the company. It notably threatened the re-authorization of the pesticide by the EU in 2022, a decision which could have led to a major shift in agricultural practices in Europe (Kudsk and Mathiassen, 2020).

Scientists debates like those about the glyphosate have gained widespread attention from researchers since the 1980s, following on the work of anthropologists of science Bruno Latour and Steve Woolgar, leading to the emergence of the field of **controversy studies** (Latour and Woolgar, 1979a). Lemieux (2007) proposes a definition of a controversy as a debate characterized by a triadic structure, where two opposing camps try to convince a real or virtual "public", and which takes place in a rather closed environment of specialists.

However, as we can see in the case of the glyphosate, some scientific debates go beyond a mere argument between specialists and involve a wide variety of actors, from activists to industrials, journalists or policy-makers. In the literature, **socio-technical controversies** have emerged as a subject of interest, with a larger meaning: this term encompasses controversies which run through society and rely on arguments of a scientific or technical nature. They typically mobilize scientific, technical, legal, moral, and political notions.

Let us illustrate with another, much studied example of a socio-technical controversy: the debate about electric cars. This debate opposes a "pro" camp composed in part of governments and pro-technology actors claiming that electric cars are less polluting and more ecological than petrol cars, and a "con" camp, made up of some ecological groups but also of conspiracy theorists (Wagener, 2024; Catellani and Lits, 2024). Ecologists notably contest the ecological aspect of these cars because of the environmental cost of manufacturing batteries. Though it hardly started as a debate between scientists, this controversy mobilizes scientific arguments, for example concerning CO2 emission rates, and technical arguments, such as those concerning the performance of the car, but also, among other, moral arguments, criticizing the inhumane working conditions surrounding the harvesting of minerals for electric batteries.

Actor Network Theory (ANT), developed by Bruno Latour, Michel Callon and

Margaret Akrich in the 1980s, is a methodology for studying socio-technical controversies (Akrich et al., 2006). This method consists in identifying the actors of a controversy and the networks which link them: the originality of ANT being that both human and non-humans, such as concepts, objects or animals, can be considered actors of the network. Much like the work of Latour and Woolgar (1979a), it puts emphasis on the impact of social features on the outcome of controversies. As socio-technical controversies have become a subject of study of their own, other methods are being used to investigate them. For instance, (Wagener, 2024) leverages discourse analysis and numerical humanities to characterize the controversy about electric cars. In the early 2000s, sociologist François Chateauraynaud developed a computer program called Prospéro (PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur) to explore public controversies (Chateauraynaud et al., 2003). This tool can be used to analyze textual data relating to controversies in order to characterize the forms of expression and argumentation used by the actors, and is intended as an aid for sociologists studying controversies characterized by a large quantity of texts. It was one of the first approaches to use computer science tools to explore socio-technical controversies, here in conjunction with sociological methods, and gave rise to the "argumentative sociology of controversy" field of study.

Meanwhile, in computer science, recent years have seen a rise in the development of **formal argumentation** as a method to study debates. In my work, I propose to investigate socio-technical controversies using formal argumentation. I have a particular focus on the **circulation of arguments**; in particular, I investigate the role of **social structures** on the circulation of arguments within these controversies. Studying these aspects from a computer science perspective, I use **social simulation** to explore the social dynamics of controversies. As such, my thesis is part of several distinct literature, which I will now introduce before delving into my research question and methods in more details.

1.1 Formal argumentation

Formal argumentation lies at the intersection of logic and computer science, involving methods to represent, analyze, and evaluate arguments using formal mathematical structures. This approach allows for the precise modeling of argumentative processes, enabling the development of algorithms that can simulate, assess, and even automate reasoning tasks. By treating arguments as formal objects, formal argumentation facilitates the modelling and understanding of debates.

One of the most influential strands of formal argumentation is abstract argumentation, introduced by Phan Minh Dung in his seminal paper on the acceptability of arguments (Dung, 1995). Abstract argumentation focuses on the relationships between arguments—specifically, how they support or attack one another—rather than on the internal structure or content of the arguments themselves. Dung’s

framework simplifies the complex nature of argumentative interactions by representing them as abstract objects. As such, debates can be represented as graphs, where nodes represent arguments and edges represent the attack relations between them. In Dung's proposition, there exists one type of relation between arguments: the attack relation. For instance, consider the following exchange of arguments: let us imagine that two persons, Greta and Carole, are having a discussion about the construction of a new highway. Carole opens with: "The new highway is necessary because it will connect a rather poor rural territory with the rest of the region." Greta answers "There is already a road and a train line, they could be modernized to serve this purpose." This discussion can be understood as being composed of two distinct arguments: a : "The new highway is necessary because it will connect a rather poor rural territory with the rest of the region." and b : "There is already a road and a train line, they could be modernized to serve this purpose.", where argument b attacks argument a . This discussion can be represented in an argumentation graph such as the one presented in Figure 1.1.

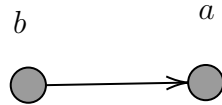


Figure 1.1: Argumentation graph representing the two arguments exchanged by Carole and Greta.

This abstraction allows for a powerful and flexible way to analyze argumentative dynamics. As such, abstract argumentation seems a promising tool to explore socio-technical controversies. However, until now, few works have attempted to study socio-technical controversies from a formal argumentative point of view.

Šešelja and Straßer (2013) apply abstract argumentation to represent scientific debates, proposing to improve abstract argumentation frameworks by adding an “explanatory” relation. This work provides valuable insights into the structure of scientific debates: for example, they show that scientific theories can be modeled by sets of arguments. We note that the authors focus their analysis exclusively on controversies taking place between scientists, without extending it to the wider set of actors that characterize socio-technical controversies. Salliou and Thomopoulos (2018) use an abstract argumentation framework to represent the debate surrounding vegetarian diets. They identify the actual arguments forming the debate and construct an explicit graph showing the attack relationships between them. This graphical representation enables them to identify the key arguments in the controversy, which occupy central positions in the graph.

These studies suggest that formalizing controversies using abstract argumentation is a way of exploring the *structure* of debates, providing information that can complement non-formal studies. Socio-technical controversies, however, are characterized by dynamic and social aspects, which cannot be represented by abstract argumentation alone. To include these aspects, abstract argumentation can be completed by **agent-based models**.

1.2 Agent-based models

Agent-based models (ABMs) are a class of models which represent individual agents, specify the interactions between them, and study the outcome of these interactions. Agent-based models implement a bottom-up approach: rules are designed at the individual agent level, rather than assuming a predetermined global structure. They are especially well suited to explore complex, non-linear phenomena such as social interactions. An important characteristic of ABMs is their ability to model heterogeneity: agents in ABMs can have different characteristics, rules, and behaviors. This allows for the modeling of diverse populations or entities within a system.

Agent-based models came from computer science, but they have a rich history that spans several disciplines: they are a multi-disciplinary tool, being used notably by economists, biologists, philosophers and sociologists. In 1971, Thomas Schelling developed one of the first agent-based models, to study racial segregation ([Schelling, 1971](#)). His model used individual agents making decisions based on local preferences, showing how even mild preferences for neighbors of the same color could lead to significant segregation. This model was instrumental in highlighting the key concept of **emergent behavior** in social systems: a model shows emergent behavior when the collective dynamics of the system are not explicitly programmed but arise from the local interactions of agents. They are well adapted to study situations of communication between humans, which consist of many local interactions, and thus, they are used in the field of **opinion dynamics** to study the evolution of agents' opinions.

Recently, agent-based models have been applied to model debates, in conjunction with the theories of formal argumentation. In particular, models such as ([Gabriellini and Torroni, 2014](#); [Butler et al., 2019](#); [Dupuis de Tarlé et al., 2022a](#)) use abstract argumentation to represent the cognition of agents, who can form opinions and exchange information using abstract arguments. This allows for the construction of enhanced opinion diffusion models to study argumentative dynamics. For instance, [Borg et al. \(2017a\)](#) represents a debate between scientists, and [Taillandier et al. \(2021\)](#) models the diffusion of vegetarian diets in a population through the diffusion of arguments in favor or against eating meat.

To the best of my knowledge, no *argumentative* agent-based model explicitly models socio-technical controversies. Agent-based models which study socio-technical controversies are part of a related literature: **computational social epistemology**.

1.3 Social epistemology

Computational philosophy is the use of computers (in particular, simulations) to inform philosophical questions. Epistemology is the study of people's beliefs and knowledge, and in particular, what forms good knowledge. Social epistemology studies how norms, structures and interactions at the group level influence the beliefs and knowledge of people.

Historically, epistemology has been studied by many philosophers such as Hume, Descartes, or Kant from an *individual* point of view. Classical epistemology focuses on an isolated individual and its means of forming beliefs and acquiring knowledge. Indeed, much of these works originated from the Enlightenment, a philosophical era which placed a heavy focus on individuals and their rational abilities. In the past 20 years, a branch of epistemology has shifted its focus on the *social* aspects of epistemology. Social epistemology recognizes that people are fundamentally part of social groups and acquire their knowledge through interactions which are determined by social structures. Thus, social epistemology studies how norms, structures and interactions at the group level influence the formation of opinions, beliefs and knowledge. Alvin I. Goldman, a precursor of the field, argued in his book *Knowledge in a Social World* that agent-based models were appropriate tools to study these questions ([Goldman, 1999](#)).

Building on his contributions, agent-based models of epistemic communities have been developed to explore these philosophical questions. These models are often similar to opinion dynamics models but include an additional feature: agents can form opinions that are epistemologically better—meaning closer to the truth—than others. By examining the "epistemic success" of agents, they can address questions like: What kinds of social networks lead to higher success rates when scientists choose between two alternatives? Do misinformation, propaganda, or biased media distort opinion formation? Kevin Zollman, Cailin O'Connor, and colleagues have explored these issues using Bayesian models, where agents investigate and update their preferences according to Bayesian probability laws ([Zollman, 2007](#); [O'Connor and Weatherall, 2018](#)). Notably, [O'Connor and Weatherall \(2019\)](#) created an agent-based model to study socio-technical controversies, incorporating not only scientists but also policymakers, propagandists, and media into the model. This allowed them to investigate how the manipulation of science shapes public opinion. These models have become highly influential in computational social epistemology, and they inspired my own models based on argumentation.

To the best of my knowledge, prior to my work there existed only one agent-based model which used abstract argumentation to represent a scientific community: the model ArgABM of [Borg et al. \(2017a\)](#). This model exclusively focuses on debates between scientists; there have not been attempts to model socio-technical controversies with abstract argumentation beyond the scientific context.

1.4 Research questions and method

The aim of my thesis is to **explore how abstract argumentation can help us understanding the dynamics of socio-technical controversies**. In particular, I'm interested in how **social structures** influence the **circulation of arguments** within these controversies. In my work, I explore this question using several approaches: both theoretical and applied, focusing either on the scientific side or on the socio-political side of the controversies.

My starting point in answering this question was to study a debate taking place between scientists. I developed two agent-based models of scientific communities, using abstract argumentation to represent their cognition and the circulation of arguments in social and public spaces. These models are innovative while being part of the tradition of formal social epistemology and, more broadly, of opinion dynamics. In particular, I investigated the role of social structures: the shape of the social network, and the presence of filters of weak arguments, on the outcome of scientific debates. The latter was used in a model which we developed to investigate the impact of cognitive bias in scientific reasoning. While developing these models, I discovered that the function that I was using to assess the strength of arguments, the *gradual semantics*, lacked certain properties necessary for an ABM simulation: this led me to propose a novel gradual semantics.

In the last part of my thesis, I studied real-life use cases of socio-technical controversies to test the representational power of abstract argumentation for socio-technical controversies. Indeed, given the lack of previous attempts, it was essential to gather empirical insights about the way these debates could be represented using abstract argumentation.

The first case of application that I studied is a debate between proponents and opponents to the construction of a large sawmill in Tarn, France. Fights against construction of GPII ("Grands Projets Inutiles et Imposés") or "Useless and Imposed Great Projects" have received a lot of media attention recently in France; they are clear cases of socio-technical controversies, although they differ from scientific-centered debates in a number of ways. I chose this case as a first test because it was representative, yet smaller and more tractable than big nation-wide debates. I built a visualisation tool and performed interviews which allowed me to represent the controversy as an abstract argumentation graph, gathering insights about the argumentative strategies of both camps.

As a second use case, I studied a well known example of socio-technical controversy: the debate about climate change. I focused on the use of scientific arguments by climate-skeptics blogs. Using abstract argumentation and qualitative analysis, I showed their internal contradictions and the accumulation of scientific "evidence" as an argumentative strategy.

Thus, my thesis can be read as a gradual move from theoretical studies towards applications: starting with insights about the properties of abstract argumentation formal tools, followed by simulation work, and ending with the study of real-life controversies. Together, these works paint a picture of controversies as dynamic processes, subject to manipulation and whose outcome is heavily influenced by social structures. This matches several conclusions from the sociology of controversies, showing the relevance of formal tools, and in particular abstract argumentation, to study these questions.

1.5 Thesis structure

This thesis is divided in three parts. The first part, "Context and Tools", presents the necessary background: Chapter 2 presents abstract argumentation in greater details and Chapter 3 agent-based models, including social epistemology. Chapter 4 propose an in-depth comparison of agent-based models which use abstract argumentation.

The second part of this thesis, entitled "Modelling Scientific Controversies with Abstract Argumentation", presents the bulk of my work: Chapter 5 introduces a general framework to model the opinion of agents with abstract argumentation, which is applied in Chapter 6, where I present an agent-based model of a scientific community. Chapter 7 is another model of a scientific community, using a different paradigm, which studies the impact of a cognitive bias on the performance of scientists.

The third part, "Exploring Applications", presents two work-in-progress which apply the results and tools developed in the previous chapters to specific use cases. In Chapter 9, I present a tool which uses abstract argumentation to understand and inform the strategies of a group of opponents to the construction of an industrial project. Chapter 10 focuses on the controversy about climate change, presenting a characterisation of the scientific arguments mobilized in climate skeptic platforms.

The work presented in my thesis has been published as: [Dupuis de Tarlé et al. \(2023\)](#) in the proceedings of the 19th Social Simulation Conference (SCC), [Dupuis de Tarlé \(2023\)](#) in the proceedings of the Journées d'Intelligence Artificielle Fondamentales (JIAF), [Dupuis de Tarlé et al. \(2024\)](#) in the Journal of Artificial Societies and Social Simulation (JASSS). [Dupuis de Tarlé et al. \(2023\)](#) received the Best Student Paper Award from the Social Simulation Conference. Some of the reviews of argumentative agent-based models presented in Part I will be published as [Dupuis de Tarlé et al. \(forthcoming\)](#) in the Handbook of Formal Argumentation Volume 3.

Part I

Context and Tools

CHAPTER 2

Abstract Argumentation

Abstract argumentation, introduced by [Dung \(1995\)](#), is a method for formalizing argumentative discussions that considers arguments as abstract objects and focuses on the relations that link these arguments together. Originally developed in symbolic AI, abstract argumentation has been increasingly used for agent-based modeling of deliberating communities (see [Chapter 4](#)).

In abstract argumentation, argumentative discussions are modelled as sets of objects and relations: as such, they can be represented as graphs. These graphs are called “argumentation frameworks” or “argumentation graphs”. The following example illustrates how abstract argumentation models an argumentative discussion.

Example 1. Suppose Ana and Pablo are arguing about the benefits of renewable energy. Ana makes the following argument, a : “Renewable energy should be preferred to fossil energy because this will help to fight global warming”. Pablo disagrees and responds with a counterargument b : “Renewable energy is more expensive than fossil one, and, as a result, it is not efficient.” Ana responds with c : “The cost we pay now to move to renewable energy is not comparable to the cost we would pay in the future if we stick with fossil fuels”. The argumentation graph capturing this debate is illustrated in [Figure 2.1](#).

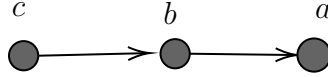


Figure 2.1: Representation of the debate of [Example 1](#) as an argumentation graph, where the arguments are represented as nodes and the attack relations as directed edges.

Definition 1. An **argumentation framework** (AF) is a pair $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ where $\mathcal{A}$ is a finite and non-empty set of (abstract) arguments, and $\mathcal{R}$ is a binary relation on $\mathcal{A}$, i.e. $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, called the attack relation. $\forall x, y \in \mathcal{A}$, $(x, y) \in \mathcal{R}$ means that x **attacks** y .

An AF may be represented by a directed graph, called the **argumentation graph** (AG), whose nodes are arguments and edges represent the attack relation.

In the remainder of this work, we will use the notion of argumentation framework or argumentation graph indistinctly. Let us now introduce different notions that we will use in this work.

Definition 2. Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework and $x, y \in \mathcal{A}$ be two arguments. A **path** P from y to x , denoted $P(y, x)$, is a sequence $\langle x_0, \dots, x_n \rangle$ of arguments such that $x_0 = x$, $x_n = y$ and $\forall i \in \{0, 1, \dots, n-1\}$, $(x_{i+1}, x_i) \in \mathcal{R}$. We

denote by $l_P = n$ the **length** of P . An argument is a **defender** (resp. **attacker**) of x if it is situated at the beginning of an even-length (resp. odd-length) path towards x .

We shall abuse notation and write $Arg(AG)$ for the set of arguments of a given AG , and $Att(x)$ for the direct attackers of a given argument x .

In this way, AFs can be used to model a stance of a rational reasoner in a debate: by assigning to each agent a specific argumentation framework we can represent a set of arguments and attacks between them, forming the agent's argumentation map. Moreover, as we shall now see, we can define the rules specifying which arguments are acceptable, or which arguments should be rejected.

2.1 Acceptability Semantics

In Dung's framework, the *acceptability of an argument* depends on its membership to some sets, called extensions. These extensions characterize collective acceptability.

Definition 3. Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework. Let $S \subseteq \mathcal{A}$. S is **conflict-free** for AF iff there exists no x, y in S such that $(x, y) \in \mathcal{R}$. S **collectively defends** an $x \in \mathcal{A}$ iff $\forall y \in \mathcal{A}$ such that $(y, x) \in \mathcal{R}$, $\exists z \in S$ such that $(z, y) \in \mathcal{R}$.

For instance, in Figure 2.2, the set $\{a, c\}$ is conflict-free. A set of arguments is *admissible* when it is conflict-free and each argument of the set is collectively defended by the set itself. Several *semantics for acceptability* have been defined in Dung (1995).

Given an AF, Dung defined different ways in which sets of arguments can be deemed acceptable in a given argumentation framework. These sets intuitively characterize rational positions in a debate. Here, we present the grounded semantics, which has the interesting property of returning only one set of arguments.

Definition 4 (Grounded set). Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF. An argument $a \in \mathcal{A}$ is **acceptable with respect to a set of arguments** E in case for every b for which $(b, a) \in \mathcal{R}$, there is a $c \in E$ for which $(c, b) \in \mathcal{R}$. The **grounded set** G for AF is iteratively defined as follows: $G = \bigcup_{i=0}^{\infty} G_i$ where G_0 is the set of all arguments which have no attackers in $\mathcal{A}$ and G_{i+1} is the set of all arguments that are acceptable with respect to G_i . We call an argument **acceptable** if it is part of the grounded set.

Figure 2.2 represents the application of the grounded semantics to the graph of Example 1. While b attacks a , c attacks b and thus defends a . Given this stage of the debate, we can say that a set of arguments $S = \{a, c\}$ is grounded and acceptable.

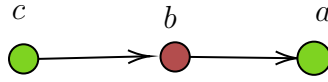


Figure 2.2: Representation of the debate of Example 1 as an argumentation graph, where the arguments are represented as nodes and the attack relations as directed edges. a , c in light green are acceptable, and b in dark red is not.

2.2 Bipolar Graphs

When they were first introduced, argumentation frameworks only contained an attack relation between arguments. Since then, the *support* relation has been introduced to allow more expressivity, giving rise to **bipolar argumentation graphs** (Cayrol and Lagasquie-Schiex, 2005). The support relation represents a positive relation between two arguments, which cannot be expressed through defense (attacking an attacker). Empirical studies have shown that bipolar argumentation graphs are more adapted to represent the way humans actually reason about arguments (Polberg and Hunter, 2018).

Figure 2.3 presents the argumentation graph which corresponds to the debate of Example 2.

Example 2. *Let us imagine that two friends, Lisa and Ana, are having a discussion about organic agriculture. Lisa opens with "All the agriculture of our country should become organic". She adds that "Non organic agriculture degrades biodiversity, soils and our health." Ana answers "But organic agriculture is less productive, we may not be able to feed the entire population." Lisa responds "We could, if we eat less meat and limit the food waste." This discussion can be understood as being composed of four distinct arguments:*

- a : "All the agriculture of our country should become organic"
- b : "Non organic agriculture degrades biodiversity, soils and our health."
- c : "Organic agriculture is less productive, we may not be able to feed the entire population."
- d : "We could, if we eat less meat and limit the food waste."

With the following relations: argument b is supporting argument a , argument c is attacking argument a and argument d is attacking argument c . Figure 2.3 represents the argumentation graph of this discussion.

To further enhance the expressive power of the graphs, a very common extension consists in equipping each argument with a "weight", which corresponds to an intrinsic quality of the argument. Weights can be assigned in a variety of ways, depending

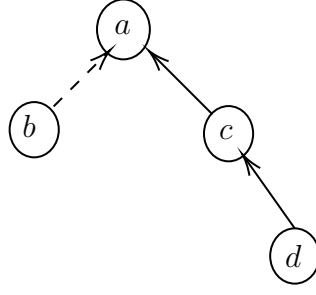


Figure 2.3: Argumentation graph representing the relations between the arguments of Example 2.

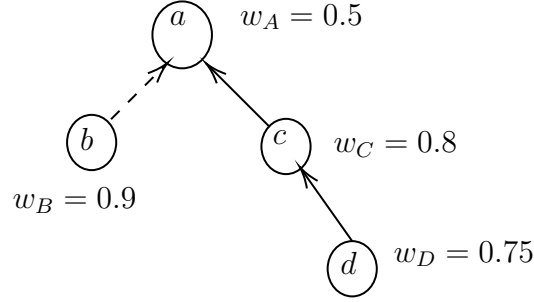


Figure 2.4: Argumentation graph of Example 2 equipped with weights for each argument.

on the modeler's choice. For instance, we could assign a weight to every argument of the discussion of Example 2, which would correspond to their persuasive power; see Figure 2.4 for an example.

Formally, we can define **weighted bipolar argumentation graphs**.

Definition 5 (Weighted Bipolar Argumentation Graphs (WBAG)). *A weighted bipolar argumentation graphs B is a quadruple $B = \langle A, R, S, W \rangle$ where A is a finite set of arguments, $R \subseteq A \times A$, $S \subseteq A \times A$ are two binary relations on arguments, respectively attack and support, and W is a function from A to $[0, 1]$.*

Note that weighted bipolar graphs include attack-only graphs and non-weighted graphs.

In order to simplify notations, we will use $B \setminus A'$ with $A' \subset A$ to denote graph B without the arguments of A' and the attack and support relations featuring these arguments. We now also define the set of attacker and supporters of an argument.

Definition 6 (Set of attackers and supporters). *Let $B = \langle A, R, S, W \rangle \in G$ a weighted bipolar graph and $a \in A$ an argument from this graph. The set of attackers and supporters of a in B are respectively defined as $Att(B, a) = \{b \in A \mid (b, a) \in R\}$ and $Supp(B, a) = \{b \in A \mid (b, a) \in S\}$.*

2.3 Gradual Semantics

A major challenge of abstract argumentation is the characterisation of the *acceptability* of arguments based on the information contained in an argumentation graph. **Gradual semantics** are functions which evaluate the acceptability of arguments through an *acceptability score*. To the best of our knowledge, all the gradual semantics for bipolar graphs proposed in the literature assign a score in $\mathbb{R}$ (Amgoud and Ben-Naim, 2018; Mossakowski and Neuhaus, 2018), and in this work we make the assumption that we always use gradual semantics whose image interval is comprised in $\mathbb{R}$.

Definition 7 (Gradual Semantic). *Let G the set of all weighted bipolar graphs, $B = \langle A, R, S, W \rangle \in G$, and D an interval of $\mathbb{R}$. A gradual semantic on B is a function $\sigma : G \times A \rightarrow D$ and for all $a \in A$, $\sigma(B, a)$ denotes the acceptability score of a in B .*

It is important to distinguish between the two numbers which characterize the arguments. *Weights* are assigned externally and represent the quality of the arguments. *Scores*, given by the semantics, use the weights and the structure of the graph to represent the degree of acceptability of the argument within the context of the discussion. Assigning weights is optional, as we can choose a uniform weight for all arguments.

We specifically focus on a particular type of gradual semantic known as **modular** semantics, as identified by Mossakowski and Neuhaus (2018). These semantics consist of two functions: one aggregates the scores of the attackers and supports of a given argument, and the other determines their influence on the base weight of this argument. Consequently, the various semantics proposed for bipolar argument graphs in existing literature can be analyzed as combinations of an aggregation function and an influence function.

Definition 8 (Modular Gradual Semantic). *Let $B = \langle A, R, S, W \rangle \in G$, and D an interval of $\mathbb{R}$. A modular gradual semantic on B is a function $\sigma : G \times A \rightarrow D$ where $\sigma(B, a) = \iota(\alpha(B, a))$ with α and ι respectively the aggregation function and influence function.*

In the rest of this work, we will only consider modular gradual semantics. This is hardly an important restriction since, as noted by Mossakowski and Neuhaus (2018), all the gradual semantics for weighted bipolar graphs presented in the literature are modular.

The Euler based semantic (EBS) introduced by Amgoud and Ben-Naim (2018), which we formally define in Section 5.3, is an example of a modular semantics.

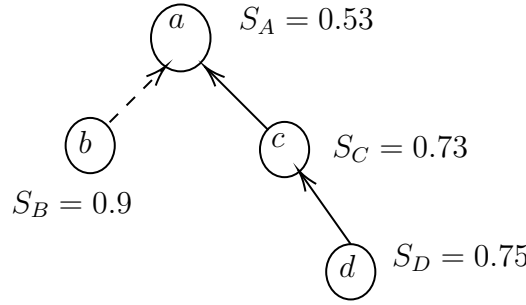


Figure 2.5: Argumentation graph of Example 2 with the acceptability scores given by the EBS semantics.

Figure 2.5 represents the score of the arguments of Example 2 when we apply the EBS semantics, using the weights of Figure 2.4. We see that argument a has an acceptability score of 0.53 in this configuration, which could be interpreted as “The proposition “All the agriculture of our country should become organic” is 53% acceptable”.

Assigning an acceptability score to arguments can be useful in many cases. For instance, an online debate platform may benefit from assigning scores to the arguments expressed by users to facilitate constructive discussions and highlight more persuasive or well-supported arguments. Another potential application is in recommender systems: gradual semantics can allow one to weight the positive and negative arguments toward an option, providing a nuanced assessment of user preferences. This would enable the system to offer better recommendations by offering a fine grained evaluation of the options, taking into account the number of supporting and attacking arguments as well as their respective strengths.

In my work, gradual semantics are used to represent the opinions of agents. This allows for a fine-grained representation of how agents process and evaluate the arguments they encounter. The possibility of using gradual semantics in these scenarios depends crucially on its mathematical properties, as developed in Chapter 5.

2.4 Empirical Validation of Abstract Argumentation

Due to its versatility and simplicity, abstract argumentation has been widely adopted by researchers, who extended and generalized Dung’s original framework. Polberg and Hunter (2018) remark that these new developments of abstract argumentation are usually justified by their authors because they create links with other abstract frameworks, or because they provide novelty or technical correctness. However, many applications of abstract argumentation use it as a way to model how humans

actually reason about arguments. This is the case of my work, where abstract argumentation is used to represent the cognition of agents (see Chapter 5). This raises the necessary question of whether this assumption is justified: do humans treat arguments the same way as abstract argumentation semantics? Several authors have performed empirical studies to test abstract argumentation. [Cerutti et al. \(2021\)](#) present a survey of these studies.

[Rahwan et al. \(2010\)](#) were the first to define the following research questions regarding the formalism of abstract argumentation:

- Do the features of the semantics represent a real cognitive feature characteristic of human reasoning?
- Which semantics most accurately represent the way humans reason?

2.4.1 Testing set-based semantics

The first empirical studies on abstract argumentation, [Rahwan et al. \(2010\)](#) and [Cramer and Guillaume \(2018, 2019\)](#) naturally focused on testing the classical, set-based argumentation semantics. These semantics (presented above) assign only two possible values to arguments: accepted, or rejected.

[Rahwan et al. \(2010\)](#) propose behavioural experiments to assess the feature of "reinstatement", which is common to all set-based semantics. The idea of reinstatement is illustrated in Example 1: according to the semantics of argumentation defined by [Dung \(1995\)](#), if an argument a is attacked by another argument b , it is rejected, unless argument b is itself attacked by an accepted argument c . In this case, argument a is defended by c , and "reinstated". Figure 2.2 represents the corresponding argumentation framework. All set-based semantics would classify a and c as accepted and b as rejected, which begs the question: would humans do the same? [Rahwan et al. \(2010\)](#) aim to test empirically whether this mechanism corresponds to real life behaviour. For this, they present to participants groups of three arguments who represent an argumentation framework similar to that of Figure 2.2. The participants rate their confidence in the argument on a 7-point scale, ranging from "Strongly Agree" to "Strongly Disagree".

The authors report the following result: that reinstatement does exist, but it is not strict. Confidence in the argument decreases when it is attacked, and increases after being defended, but does not reach the same level as before the attack. We could call this "partial reinstatement". For instance, a participant may have expressed a degree of confidence of "Strongly Agree" for argument a , and when confronted with arguments b and c , would report that they "Somewhat Agree" with the argument.

One way of interpreting this results is as supporting the use of fine-grained approaches to describe the acceptability of arguments: according to this study, classical, two or three-values semantics would be insufficient to describe how people actually reason. On the contrary, gradual semantics, who can assign degrees of acceptability in the interval $[0, 1]$ are apt to describe such situations: for instance, the novel semantics that I propose in Chapter 5 models partial reinstatement.

This result is mitigated by further experiments. Cramer and Guillaume (2018, 2019) also investigated reinstatement, building on the methodology of Rahwan et al. (2010). In their two studies, however, they limit the answers of the participants to three possibilities: accepted, rejected or undecided. Thus, their method is less suited to discriminate between fine-grained approaches and classical semantics. They find that the participants follow the principle of reinstatement: this suggests that set-based semantics do represent human reasoning in some measure. However, the question of whether a wider range of answers would have shown partial rather than total reinstatement remains open.

2.4.2 Testing fine-grained approaches

Polberg and Hunter (2018) perform a study in which participants are asked to determine the attack and support relations between arguments about the flu vaccine. They are also asked to judge their acceptability on a 7 point scale. The authors aim to investigate the descriptive power of **probabilistic argumentation**.

Probabilistic argumentation is an extension of abstract argumentation developed by Hunter (2013) which models *uncertainty* concerning abstract argumentation graphs and the acceptability of arguments. It consists of two main approaches.

1. The **constellation** approach represents situations where we are unsure about the *relations* between the arguments. This is represented by uncertainty in the **topology** of the argumentation graph. Formally, it is modelled by a probability distribution over the space of possible argument graphs, where each graph has a chance of being the "real" one.
2. The **epistemic** approach represents uncertainty in the belief of the agents about argument acceptability. Every argument is equipped with a probability. The probability of a single argument is understood as the belief that an agent has in it, i.e. we say that an agent believes an argument A to some degree when $P(A) > 0.5$, disbelieves an argument to some degree when $P(A) < 0.5$, and neither believes nor disbelieves an argument when $P(A) = 0.5$.

As such, the epistemic approach describes the acceptability of arguments in a fine-grained way, like gradual semantics; in Section 2.5, I will provide a comparison between the two approaches.

Polberg and Hunter (2018) report that the constellation approach is supported by the results of the study, as the participants do not all agree on the shape of the argumentation graph. They also note that **three values are not sufficient** to describe the participant’s opinions. They interpret this as supporting the epistemic approach to probabilistic argumentation, but this observation more generally supports the relevance of fine-grained approaches. Thus, this result can be interpreted as supporting the use of gradual semantics as well.

In the study, participants identify positive links between arguments. The authors show that some, but not all of these positive links can be modelled as defense. The fact that there are positive links which are not defense prompts their observation that **bipolar argumentation frameworks are more representative of the way people actually reason**.

Vesic et al. (2022) perform experiments to test whether the principles of gradual semantics accurately represent how people reason. They also test how showing the graphical representation of a debate influences participants. They use attack-only graphs. The participants are provided with the text arguments, and the graph representation if they are in the graph group. Then, they are asked to estimate a level of strength to all arguments by using a 4-point Lickart scale. Note that, unlike in Polberg and Hunter (2018), the participants are not asked to elicitate the relations between arguments. They are either given the relations or not given the relations, and are then asked to determine the acceptability of arguments.

The principles of gradual argumentation semantics are properties which can be verified by gradual argumentation semantics: see Chapter 5 and in particular Table 5.1 for principles of gradual semantics for weighted bipolar graphs. For instance, Anonymity is a principle which states that the score of an argument is independent from its identity, that is, two arguments placed in a similar graph have the same score.

Unsurprisingly, the authors find that participants who are given the graph representation are more likely to verify the principles of gradual argumentation semantics. The principle of Maximality, which claims that non-attacked arguments should have the maximal acceptability value, is not well verified in both groups. Vesic et al. (2022) hypothesize that this principle is too strong, and that it is not justified because of participant have background knowledge about the subjects mentioned by the arguments. The participants use this background knowledge to judge arguments regardless even when there are no attack relations. This justifies the use of initial weights to represent an initial opinion on arguments regardless of their position within the graph. Thus, the authors suggest using weighted argumentation graphs.

Empirical studies tend to favor the most general approaches: weighted graphs rather than non-weighted graphs, bipolar rather than attack; and fine-grained rather than set-based. These methods seem to be better at representing the complexity of human cognition. The following section will review the respective advantages of

different abstract argumentation approaches.

2.5 Comparing Abstract Argumentation Approaches

As we have seen, several distinct approaches generalize and extend Dung’s argumentation frameworks: bipolar graphs, weighted graphs, probabilistic argumentation and gradual semantics, among others. This section will compare these different approaches to abstract argumentation, citing their pros and cons, and justifying which ones I chose for the work presented in the following chapters.

2.5.1 Fine grained or set-based semantics

Set based semantics can assign two, sometimes three values to the arguments: accepted, rejected or undecided. On the contrary, fine grained approaches can assign many different values to the acceptability of an argument. In particular, gradual semantics and the epistemic approach to probabilistic argumentation define degrees of acceptability in the interval $[0, 1]$.

I used both approaches in my multi-agent models: Chapter 7 uses a set-based semantics, Chapter 6 uses a gradual semantics. The main advantage of using a gradual semantics is that we can define the opinion of agents as a number between 0 and 1. This is similar to a lot of opinion dynamics model (see the following section). Using a gradual semantics allowed me to compare my results with these models, and to borrow some of their methods. But, because there were not many options of gradual semantics for bipolar graphs, I had to create a new one which verified the necessary principles (presented in Chapter 5). On the contrary, the use of a set-based semantics in the model of Chapter 7 led to the construction of a complex protocol, to allow for a more fine-grained evaluation of the agent’s beliefs.

The question of which set-based semantics is the most appropriate is much discussed in the abstract argumentation literature. For instance, [Cramer and Guillaume \(2019\)](#) aim to answer it by comparing the participants’ answers to that of several set-based semantics. However, this question was not relevant for my work: these semantics only differ in argumentation graphs that feature cycles ([Dung, 1995](#)). In my work, all the graphs were cycle-free. This is why in the model of Chapter 7 where we used a set-based semantics, we did not need to choose between the classical semantics, as any of them would give the same result.

2.5.2 Bipolar or attack graph

One of the first questions when building a model based on abstract argumentation is the type of relation that we wish to consider. Dung’s original frameworks only contain attacks. The majority of semantics, in particular set-based semantics, have been developed for attack graphs. Bipolar graphs were introduced by [Cayrol and Lagasquie-Schiex \(2005\)](#) as a generalization, by adding the support relation. The support relation is a positive relation between two arguments which cannot be described by a defense, i.e. the attack of an attacker.

The agent-based model developed in Chapter 7 uses attack graphs, while the model of Chapter 6 uses bipolar graphs. I chose to use bipolar graphs in this approach because I wanted to build a protocol where the opinion of the agents could increase or decrease, and using support allowed me to build a symmetrical protocol - if I used attack-only graphs, the opinion was more likely to decrease, as it was easier to generate an attack than a defense (for which we need two arguments). The use of attack graphs in Chapter 7 led us to create a simpler protocol for argument generation. The problem of symmetry was not present because we did not use gradual semantics. Having experimented with both, I would say that when creating agent-based models, bipolar graphs allow for more creativity and versatility, but that attack graphs are a sufficient option in many cases.

As we have seen in the previous section, [Polberg and Hunter \(2018\)](#)’s study supports the use of bipolar graphs as being closer to how humans reason. When performing the case study of Chapter 9, I also found bipolar graphs very useful to describe the situation that I was studying. Chapter 10 also highlights the usefulness of the support relation, which allows us to transform an existing taxonomy of climate skeptic claims into an argumentation framework.

2.5.3 Gradual or epistemic approach

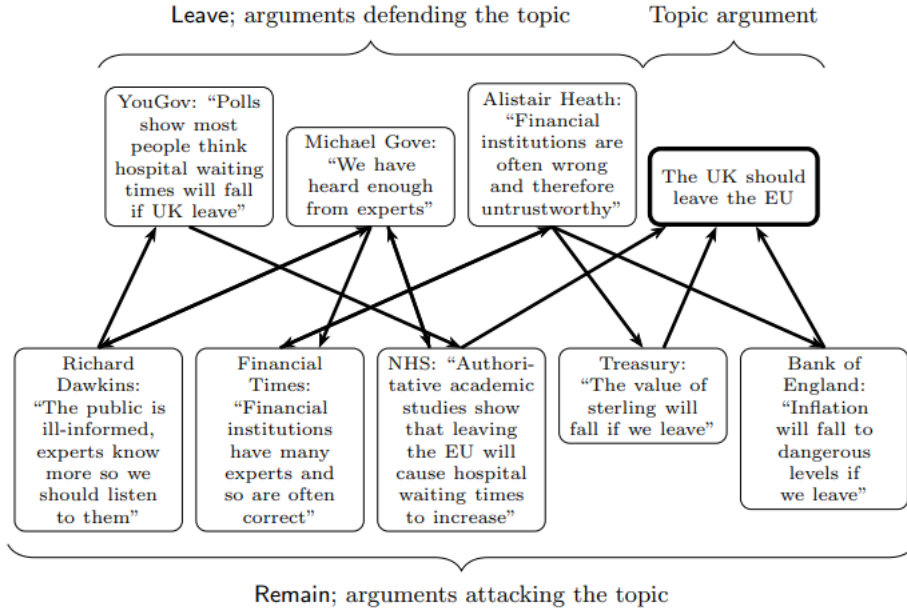
The epistemic approach to probabilistic argumentation (EA) and the gradual approach have the same advantage: they are more expressive than set-based semantics. EA represents the **uncertainty** of the agent. The gradual semantics do not express uncertainty, but a fuzzy degree of acceptability: note that in my work of Chapter 6, the scores of the arguments are compared to truth values which can be understood as fuzzy truth values. The difference between probabilistic argumentation and gradual semantics are similar to the difference between probabilities and fuzzy logic. To illustrate the difference between a probability and a fuzzy truth value, consider the following example: using probability theory, we can say that Paul has a 70% chance of being bald, which means that he is either fully bald or not at all, but we don’t know for sure and can only express a degree of uncertainty. Fuzzy logic can allow us to say that Paul *is* 70% bald: this means that we are certain that he has only about three quarters of his hair.

The difference between epistemic and gradual approaches is similar: in the former, one considers that arguments are either accepted or rejected but that agents may be uncertain about it, while the latter considers that the degree of acceptability of the argument itself is non binary. For instance, an agent could have a degree of belief of 70% in the argument "The science is used by climate skeptics is false". In EA, the argument is either acceptable or it is not, ie, given the information available, the agent could fully assert that "The science used by climate skeptics is false", meaning all the science is false, or it could not and the argument would be rejected. The degree of belief indicates that the agent is 70% sure that the argument is completely acceptable. Conversely, if an agent using a gradual semantics assigns an acceptability score of 70% to the argument "The science used by climate skeptics is false", this would mean that given the information of the graph, the argument *is* 70% acceptable to this agent, which could indicate, for instance, that about 70% of the scientific articles cited by climate skeptics are false. In short, EA is about uncertainty and beliefs while gradual semantics can express multiple levels of acceptability.

Gradual semantics have the advantage of being easy to implement. Many gradual semantics are defined by a function which verifies a number of mathematical properties corresponding to "principles". This makes gradual semantics well fitted to use in multi-agent models. Conversely, probabilistic argumentation necessitates the implementation of many complex logical principles. To the best of my knowledge, it has never been used in an agent-based model.

2.5.4 Weighted or non weighted graphs

Weighted graphs, where each argument is characterized by a numerical weight expressing its intrinsic strength, are more expressive and supported by empirical experiments (Vesic et al., 2022). Again, I used both methods in my work: the models of Chapter 6 and 7 use non-weighted graphs, while Chapter 9 features a weighted graph. When building multi-agent models, simplicity is desirable: as I develop in the following section, the popular KISS - "Keep it Simple, Stupid!" approach says that a model should contain only the elements that are necessary to investigate its research question. Given the protocols of Chapter 6 and 7, weights would not have been very meaningful nor would have changed the results. Thus, using non-weighted graphs was a useful simplification. When working on the case study of Chapter 9, I realized that weights were necessary to represent the situation adequately. Indeed, I noticed that the arguments were of very different "strength", and that if I did not include weights, the result would be quite unrepresentative. I could not treat all arguments the same way. Like in the empirical studies presented above, confrontation to real-life supports a more complex approach to abstract argumentation.

Figure 2.6: Example of bipartite graph representing the debate about the Brexit referendum.

2.6 Multi-Agent Protocols

In order to explore, compare, and assess various formal tools of abstract argumentation, researchers have designed protocols involving multiple agents. These protocols allow for the study of how debates evolve in dynamic environments. This section presents two influential studies: [Marzagão et al. \(2018\)](#) and [Bonzon and Maudet \(2012\)](#). Although these models do not involve simulations and are therefore not technically agent-based models (which will be described in Chapter 3), they offer valuable insights into the dynamic aspects of debates. A key distinction between these models and agent-based models using abstract argumentation (discussed in Chapter 4) is that the positions of the agents remain fixed throughout the debate. These protocols focus on situations of *persuasion*, where actors with predetermined positions seek to influence a deliberative discussion.

2.6.1 Team persuasion

[Marzagão et al. \(2018\)](#) present a model where two teams of agents debate over a certain topic. This modelling was inspired by the Brexit campaign in the UK: the agents represent politicians and institutional actors who exchange arguments in order to convince the general public to vote "Yes" or "No".

The arguments composing the debate are modelled as a bipartite graph where "pro" and "con" arguments either attack or defend the central topic. Figure 2.6

presents an example of such a bipartite graph around the topic of the Brexit referendum: the topic argument is "The UK should leave the EU".

In the model, the agents are each equipped with one argument, and at each turn they choose whether they assert it publicly (and it becomes visible to the audience) or not. For instance, in the example of Figure 2.6, an agent may represent the NHS and assert the corresponding argument "Authoritative academic studies how leaving the EU will cause hospital waiting times to increase".

The **audience** has no memory and thus it accepts the topic argument if the topic argument is acceptable given the currently publicly asserted arguments. The audience and the agents use the same acceptability semantics, which is not specified by the authors. The agents may pursue different strategies which change their decision. We refer the reader to the original article for an in-depth description of these strategies.

Marzagão et al. (2018) define **state-stable configurations**, which corresponds to states of the game where no change in the assertion of arguments can occur again. This is a state where the debate is "over" and the audience cannot change their mind anymore. The authors show analytically that the probability of either team winning can be computed, but that there are **many games which do not become state-stable**. This means that for many instances of the game, the outcome of the debate -that is, the choice of the audience- may change if we allow the game to continue. There is no definitive, stable outcome. The Brexit referendum took place at a certain date in time: the model suggests that this may have influenced the outcome and ultimate choice of the population. It is possible that if the referendum had taken place a few days later, the outcome would have been different.

2.6.2 Multiparty persuasion

Bonzon and Maudet (2012) also model a debate where the agents are divided into two teams, respectively PRO and CON a topic argument which they call the *issue*. The agents perform moves in order to change the acceptability of the issue in the public gameboard.

The particularity of this work is that agents all know about the same arguments but disagree about the attack relations between them. Figure 2.7 presents an example of such a situation: agents a_1 , a_2 and a_3 are aware of arguments a, b and the issue c but each of them considers different relations between them. By applying the *grounded* acceptability semantics, we obtain a set of acceptable arguments for each agent, here denoted $\mathcal{E}(AS_i)$. The position of the agents depends on whether the issue is part of this set of acceptable arguments. Here, only a_2 deems the issue c acceptable and is thus part of the PRO team, while a_1 and a_3 are CON.

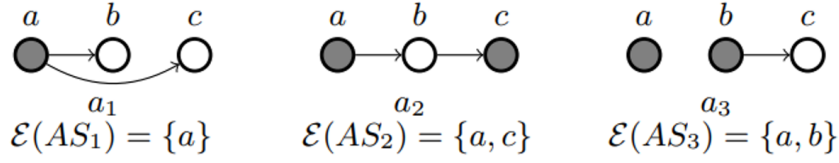


Figure 2.7: Example of three agents a_1 , a_2 and a_3 and their respective argumentation graphs.

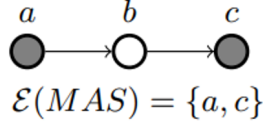


Figure 2.8: Merged graph of the agents a_1 , a_2 and a_3 .

Protocol Agents play one by one, by either asserting attack relations between arguments or opposing them. They can assert the relations contained in their individual graphs, and conversely oppose relations which do not exist in their graph. For example, agent a_1 may assert the attack relation (a, b) , or oppose relation (b, c) . This results in a **public debate** which is an argumentation graph containing the attack relation which are asserted by a strict majority of agents.

The **outcome** of the debate is obtained by applying the grounded semantics to the public debate. Figure 2.9 presents an example of a sequence of moves allowed by the protocol, and the corresponding public graphs. The outcome is that the issue is not acceptable, as we see it is not in the grounded extension of the public graph presented on the last line.

The authors are interested in how the outcome of the debate compares with the status of the issue in **the individual preferences of the agents**, as well as in the **merged graph**.

Merged Graph The merged graph is an argumentation graph obtained by merging the individual graphs of the agents. An attack relation between two arguments is kept in the merged graph if it is supported by (it belongs to the graph of) a strict majority of agents. Figure 2.8 presents the merged argumentation graph which corresponds to the example of Figure 2.7.

Note that the procedure to obtain the merged graph is very similar to the one used to obtain the outcome of the debate: only relations supported by a strict majority of agents are kept. The only difference is that the merged graph is static, while the public graph is the result of a dynamic process where the agents strategically assert or oppose relations, playing one after the other. One of the research questions of this study is whether this dynamic process will result in a different outcome than the static aggregation.

A sequence of moves allowed by the protocol is the following:

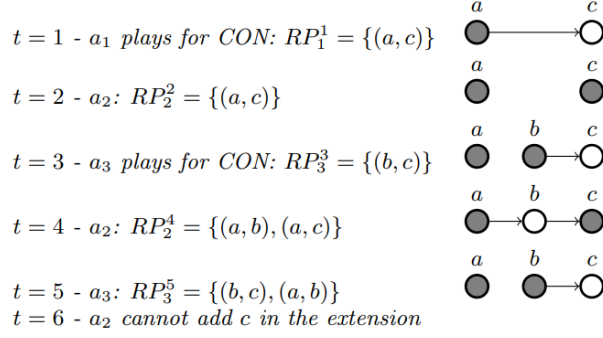


Figure 2.9: A sequence of move allowed by the protocol.

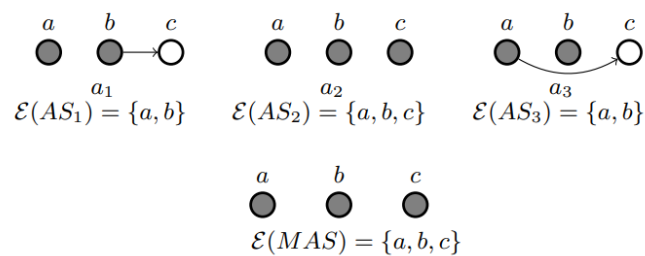
Outcome and merged graph In the example of Figure 2.8, the issue is acceptable in the merged graph, while we see that in Figure 2.9, it is not acceptable in the final outcome. Thus, the outcome and the merged graph do not necessarily coincide: this means that the dynamic aspect of the debate is not trivial, and results in situations which cannot be derived from the initial distribution.

Furthermore, [Bonzon and Maudet \(2012\)](#) also show that the outcome of a debate is not pre-determined: from a given initial situation, the issue may end up being acceptable or not, depending on the moves of the agents, some of which are randomly chosen amongst several possibilities. This makes the outcome *contingent*, in the sense that the final outcome could be different even in similar circumstances.

Preferences of agents In the example presented in Figure 2.7 and 2.8, a majority of agents are part of the CON team, and yet the issue is acceptable in the merged outcome. Figure 2.10 presents a debate with only one possible outcome, in which the status of the issue differs from the preference of the majority of the agents. Thus, both in the merged graph and the common graph, the status of the issue can contradict the individual preferences of the agents.

2.6.3 Conclusion

The results from [Marzagão et al. \(2018\)](#) and [Bonzon and Maudet \(2012\)](#) show that the outcome of a debate is highly *contingent*: it can depend on factors such as when the debate concludes, the dynamics at play—i.e., the sequence in which agents participate—and stochastic elements. Moreover, the outcome may not always align with the individual preferences of the agents involved. This highlights that the *dynamics* of argumentative discussions are far from trivial; instead, they play a crucial role in shaping the result of a debate. These findings emphasize the importance of using **agent-based methods** to explore and analyze argumentative discussions.



The issue of the dialogue is the argument c . We have $CON = \{a_1, a_3\}$, $PRO = \{a_2\}$. Agents in CON can attack c in two ways: either a_1 can play bRc ; or a_3 can play aRc . But a_2 will be able to remove either attack, and CON agents will not have the possibility to counter-attack. We will obtain $\mathcal{E}(AS(GB)) = \{a, b, c\}$.

Figure 2.10: Example of a debate with only one possible outcome, in which the status of the issue differs from the preference of the majority of the agents.

CHAPTER 3

Agent-Based Models

In this chapter, we provide an overview of agent-based modelling, including a description of the most influential strands of works which are connected to the subsequent parts of this dissertation: opinion dynamics and computational social epistemology. The first section describes general concepts of agent-based modelling. The second section gives an overview of the field of opinion diffusion. The third section presents agent-based models of scientific communities. Some of the description of the models have been adapted from the chapter "Agent Based Model of Argumentation" of the upcoming Handbook of Computational Argumentation Volume 3: [Dupuis de Tarlé et al.](#) (forthcoming).

3.1 Generalities on Agent-Based Models

Agent-based models (ABMs) are computational¹ models where individual entities, the "agents", are equipped with certain characteristics and behaviors and made to interact together in a given simulated environment. Agent-based models simulate the use of *individual* agents to study emergent phenomena at the *global* level. They can be used in many contexts: agents can represent animals, cars, drones, people... They provide especially useful insights in the social sciences, as they can represent complex social behaviors and emergent phenomena.

The notion of **emergence** is central to agent based modelling. Emergent phenomena are phenomena observed at the macro level which could not be predicted or intuited by knowing the behavior of the agents at the individual level. This concept became of importance in the 1990's with the use of cellular automaton such as Conway's Game of Life ([Gardner, 1970](#)). The Game of Life consists of an infinite two dimensional grid whose cells can be either "alive" or "dead". The cells may change their state at each time step by obeying very simple rules which only depend on the state of their neighbors. By changing the initial state of the cells, one can obtain rather complex dynamics, in particular situations where patterns of cells "reproduce" themselves. These situations occur even though no central intelligence controls the game: this is what has prompted analogies with living organisms. Expanding on this, Wolfram devotes his 2003 book "A New Kind of Science" ([Wolfram and Gad-el Hak, 2003](#)) to the exploration of cellular automata. He performs an in-depth study

¹Agent based models can be implemented within most programming languages. In practice, most implementation use either Python, which has many specific libraries for ABMs like Mesa, or dedicated platforms such as NetLogo or Gamma.

of many types of rules, showing that some result in completely unpredictable behavior. This idea that simple micro phenomenon can result in complex and surprising macro phenomenon is the basis of agent based simulation.

The notion of emergence is illustrated by a very famous example of ABM: Schelling's model of segregation (Schelling, 1971). Schelling created his 1971 seminal model in order to explain the patterns of racial segregation observed in American cities. In this model, agents are placed on a two-dimensional square grid, and can "move" randomly to one of the cells if this cell is empty. This is a very simple simulation of the housing market. The agents move only if they are dissatisfied with their current neighbors: indeed, the agents are one of two colors (red or blue), and prefer to be surrounded by a certain percentage of agents of the same color as them. If there are less than this percentage, they move. As is shown in Figure 3.1, the surprising result of this model is that segregated areas (areas dominated exclusively by one color) emerge when the agents have a social preference as low as 30% of similar others. Schelling concludes from this finding that even a small preference for similarity among neighbors can trigger a massive effect. This exemplifies weak emergence: agents collectively generate an unexpected social dynamic that diverges from their individual intention.

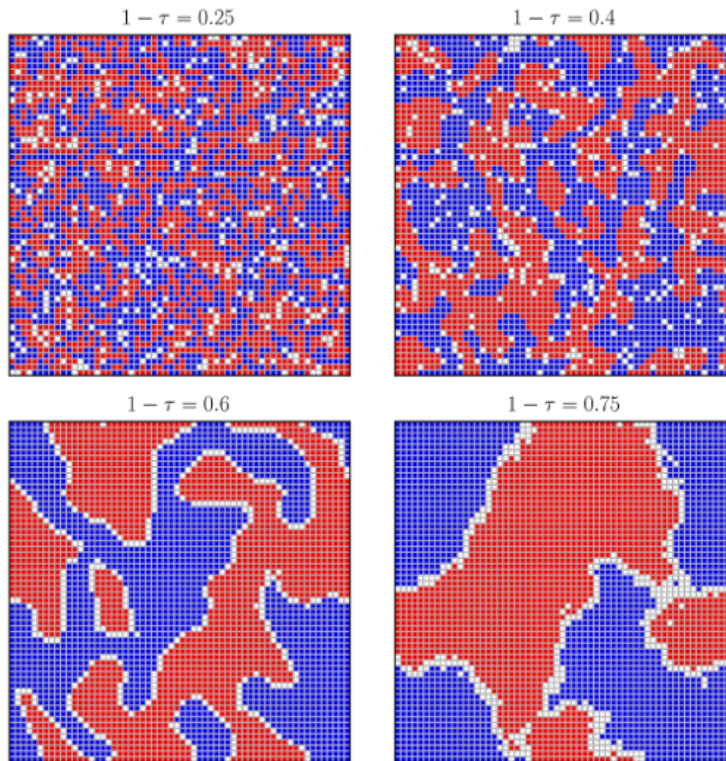


Figure 3.1: Outputs of Schelling's model of racial segregation for different levels of social preference. $1 - \tau$ is the minimal proportion of agents of a similar color that agents consider necessary.

As we see from this example, creating a model implies simplifications and idealized assumptions. In the case of the Schelling model, modelling the exact dynamics of the housing market is not necessary to make the point, would be very time and

resource consuming and ultimately, impossible. Thus, one of the first questions one may ask when creating a model is: what does the model include? How detailed should it be for its specific purpose?

3.1.1 Building a model

Two alternative approaches to the creation of a model have been proposed in the agent based modeling community, under the acronyms KISS and KIDS.

KISS: Keep It Simple, Stupid! The majority of simulations in the social sciences follow the KISS approach, which stands for "Keep It Simple, Stupid!" ([Axelrod, 1997a](#)). It consists in keeping the model rather abstract, by modeling only the relevant features in the phenomenon studied. These models allow us to understand broad mechanisms which influence general social phenomena, e.g. "segregation in cities", but cannot provide realistic predictions of the world. Typically, models start out by being very simple, and can be progressively updated with additional mechanisms once the basic dynamics of the baseline case are known.

KIDS: Keep it Descriptive, Stupid! There exists an alternative 'KIDS' ("Keep it Descriptive, Stupid") approach, introduced by [Edmonds and Moss \(2004\)](#). This approach advocates the creation of realistic, descriptive models: these models aim to describe more local phenomena and to be validated with empirical studies. Simplification is only applied when it is justified.

In this work, in line with the tradition of ABMs in the social sciences, I apply the KISS methodology to test the influence of simple mechanisms. The models presented in Chapter 6 and 7 only contain the elements that were necessary to explore their respective research questions. This raises the question of what we *can* learn from such simplified models. Does the model of Schelling truly explain what is going on in racially segregated cities? Can it prove anything about how people actually behave? These questions have been a subject of discussion in the community for decades (see ([Squazzoni, 2009](#))). Here, I provide a general overview of the main aspects which have been theorized.

3.1.2 Epistemology of ABMs

Conclusive results [Thiriot \(2020\)](#) describes the criteria which make the results of a model **conclusive**. Simulation can be thought of as a form of projection between the space of parameters towards the space of dynamics. Useful models restrict this space, otherwise, nothing can be said: this is called the "filter" effect.

Results which show a filter effect are called conclusive. He illustrates this with the example of a model which explores the spread of an epidemics given different social networks: if the model shows that any outcome is possible (from the end of the epidemics to everyone getting infected), then its results are not useful, for instance for a policymaker trying to plan a strategy.

Models as thought experiments Mayo-Wilson and Zollman (2021) argue that agent based models may serve the purpose of enhanced thought experiments. Thought experiments are theoretical, abstract situations that are used by philosophers to illustrate certain general principles, possibilities or impossibilities. As noted in Rouchier (2024), agent-based models, by leveraging the computing power of simulations, make it possible to represent much more complex situations than classical thought experiments. This makes them particularly well suited to the exploration of the dynamics of social systems where the number of agents and interactions is high.

Models as proof of concepts Abstract, idealized models can have value by showing a "proof of concept": that is, the theoretical possibility that a certain mechanism is responsible for a specific outcome. For instance, Mäs and Flache (2013) show that the exchange of arguments can be responsible for opinion polarization.

Models as "how-possibly" explanations Šešelja (2022) distinguishes between two types of explanation which can be offered by models: How-Actually Explanations, which detail how phenomena actually unfold, and How-Possibly Explanations (HPEs), which highlight how a phenomena *could* arise from certain initial conditions.

In these contexts, the results of theoretical models do not need to be compared to real-life data to be meaningful. But then, how do we guarantee the quality of the results of a model?

3.1.3 Verification and validation of ABMs

Verification and reproduction of results Simulation models are sensitive to changes in initial assumptions or conditions. To ensure transparency and reproducibility of results, the community has put in place strict model-sharing methods. A model-sharing site called CoMSES.net was created in 2008, where researchers submit their models along with a standardized description. This allows readers to check whether the assumptions are correctly reflected in the model and whether the results are consistent. Since 2010, leading journals such as JASSS (Journal of Artificial Societies and Social Simulations, the reference international journal of

agent-based modeling) have been asking authors to provide their source code to allow reviewers to verify their claims.

Another way to verify the results of a model is to perform **robustness checks**. This means changing the parameters of the model to see whether results hold. [Hales et al. \(2003\)](#) argue that tests of robustness and verification of results are facilitated when researchers adopt a "Model-to-Model" approach, which involves comparing different ABMs with each other: for instance, comparing different models that announce the same type of results and trying to see if they are actually aligned. Similarly, [Aydinonat et al. \(2021\)](#) defend for what they call the "Family of models" perspective as an even stronger test of robustness. The idea is to use different methods to model the same phenomenon, and to compare the results of these different approaches. This makes it possible to check whether certain results are only linked to some specific modelling choice. For instance, in the ABMs presented in Chapter 6 and 7, I present models of scientific community which use abstract argumentation, which is a completely different framework than previous models based on Bayesian networks.

Validating a model Models which are used as thought experiments or proofs of concepts do not need empirical validation ([Mayo-Wilson and Zollman, 2021](#)). But validating a model with data from the real world is very valuable, and [Edmonds and Moss \(2004\)](#) argue that more models should consider some form of data validation. This is especially well suited to more local models who represent specific geographic situations (e.g. a model of road traffic in a certain neighborhood). For instance, [Xu et al. \(2020\)](#) created a model of transition from conventional to organic farming, using data from French agricultural census to initialize the farmer's population. Another way of validating a model with empirical data is to *calibrate* the model: to specify the values of the parameters of the model in order to produce an output which fits with known empirical data.

Using models ethically In ABMs, only algorithms are manipulated: thus, unlike human or animal experimentation, the ethics of ABMs have been relatively unexplored until recently. This is changing, and as modelers are increasingly asked to provide input for policy-making and public decisions, the question of ethics is becoming of interest to the community, as is explored in [Rouchier \(2024\)](#). Minimal ethical criteria, related to the verification and replication of the results obtained with ABMs, already exist within the community. On the contrary, since the use of ABMs to inform large-scale public decision is relatively new - a major example being the use of ABMs to justify lockdown measures at the beginning of the Covid-19 crisis ([Rouchier, 2024](#)) - discussion around the ethical ramifications of these practices are starting to emerge. Overall, there is a consensus that using models to inform policy-makers and make predictions without analysis or transparency is a bad practice: indeed, many results of ABMs are dependant on initial hypotheses and parameters (such as the shape of the social network ([Thiriot, 2020](#))), and are

hard to calibrate and validate, even more so when facing an unprecedented situation for which we have no available data.

Some researchers have developed ways of using these models as a tool for dialogue, oriented towards solving collective decision problems. Researchers and participants build a toy model together, and the process of creating the model, checking its output and refining the hypotheses is a way for the actors to improve their understanding of the situation. ([Barreteau et al., 2013](#)).

3.2 Opinion Dynamics

Opinion dynamics is an area of research located at the intersection between computational philosophy and social epistemology. It is concerned with the question of how beliefs and opinions evolve within a social group, under different circumstances. The main assumption is that people modify their opinions, attitudes, behaviour or beliefs based on social interactions ([Akers et al., 1995](#)). It uses agent-based model simulations to explore the complex interplay between the individual and the social processes on the dynamics of the opinions. Here, we use opinion as a generic term which may include agent's beliefs, judgements or positions regarding an issue. Several sociologists have argued that individual opinion does not exist in itself: in his chapter "L'opinion publique n'existe pas" (Public opinion does not exist), [Bourdieu \(1973\)](#) criticizes the concept of public opinion obtained through opinion polls, stating that it is mostly a political tool. In "Appartenances multiples, opinion plurielle", [Ferrand \(2016\)](#) describes situations where the same people defend opposite points of view in different situations. He proposes a definition of opinion as something formed by the structure of interaction between people and not existing as predefined in people by themselves. Thus, the question of how to define opinion is rather complex. Here, we will adopt a modeler's point of view: the opinion of an agent is a position in relation to an issue that is subject to the influence of others. Opinion is what has regularity as a function of influence and aggregation, and as modellers, we are interested in what a change in framework induces in terms of change in this regularity.

Opinion dynamics, which is sometimes referred to as opinion diffusion, typically tries to provide explanations for the global trends and dynamics of opinions that can be observed in the real-world, for instance through opinion polls, or to study the influence of a known social factor in the dynamics of opinion. Thus, a certain number of questions are common to many opinion dynamics model. We start by presenting a few of these important concepts, such as polarization, network effects or cognitive biases. Then, we present classical and influential models of opinion dynamics, before exploring how argumentation can be used in opinion dynamics models.

3.2.1 Key concepts

Polarization and homophily Polarization is a central question in many opinion diffusion models. The term "polarization" describes a phenomenon where public opinion becomes more and more fragmented into two (or more) distinct groups. As such, polarization is a distinct outcome from consensus, where all agents share a similar opinion, and dispersion, which corresponds to agents who hold a wide range of opinions. There is a large literature on the progressive polarization of American society, as noted by Bramson et al. in "Understanding Polarisation: Meanings, Measures, and Model Evaluation" (Bramson et al., 2017a)². Polarization is a subject of great interest in the opinion diffusion community, in part because the first mathematical models of opinion were incapable of producing any form of polarization (Abelson, 1964). This created an interest in the type of social mechanism that could explain the polarization which was observed in practice. Bramson et al. (2017a) remark that there is no precise and unique definition of polarization. They distinguish nine phenomena that may be considered polarization. For instance, "spread" is defined as the difference between the two extreme opinions, whereas "dispersion" is a statistical measures of the dispersion of opinion, and both could be defined as polarization.

Homophily is a concept in sociology which describes the tendency of individuals to interact preferably with similar others. It is commonly represented and studied in agent based models of opinion, especially as a potential factor driving polarization. Homophily can be modeled either as agents having a higher probability of interacting with those who have a similar opinion, which we call "likelihood-based homophily" (Dupuis de Tarlé et al., [ming](#)), or as agents having more influence on each other when their opinions are more similar, which we would call "effectiveness-based homophily".

Network effects Opinion diffusion models also investigate how certain types of social structures influence the evolution of opinions. Usually, this is represented through **social networks**: graphs whose nodes are the agents and edges are social links between them. In models which use social networks, agents typically only interact with their neighbors in the graph. Modelers may compare the behavior of agents in various social network structures. For instance, Zollman (2007) compares the output of his model in four types of social networks: the cycle, the wheel and the star, as shown in Figure 3.8.

In order to create the social networks of agents, modelers often use random graphs generators: programs which take certain parameters as input and return a random graph with specific properties. One of the most simple and commonly used are Erdős-Renyi random graphs generators (Erdős et al., 1960). An Erdős-Renyi

²The authors also remark that the apparent polarisation of public opinion in the U.S. can be put into perspective given that the disagreement on a number of once very polarizing issues (e.g., gay marriage) has decreased consistently over the last decades in all social groups.

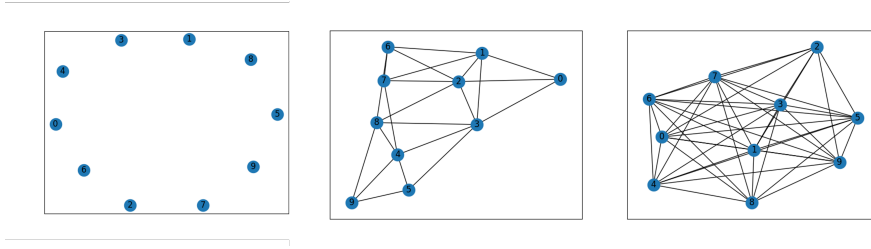


Figure 3.2: Examples of Erdős-Renyi random graphs with $n = 10$ nodes and the probability of connexion p being, from left to right: $p = 0$, $p = 0.5$, $p = 0.98$.

graph generator takes as parameters the number of nodes n and the probability of connection p . An Erdős-Renyi random graph is constructed by connecting the n nodes randomly, each edge being included in the graph with probability p , independently from every other edge. This social network does not enforce any particular social structure, and the parameter p allows the modeller to control the density of the graph; as p increases from 0 to 1, agents are more and more likely to be connected. Figure 3.2 presents three examples of networks obtained from the Erdős-Renyi graph generator with parameters $n = 10$, and different values of p .

Small-world networks are a class of graphs which are particularly useful to model social structures. They are characterized by a small shortest-path distance between nodes (the distance from an agent to one of its neighbors is 1, the distance from an agent to a neighbor of a neighbor who is not also its neighbor is 2, and so on). In a small world network, the average distance between any two agents is small: usually, less than 10. These characteristics have been identified in many real-life networks, thus small-world networks are usually considered good approximations of real social networks. Thiriot (2020) identifies five kinds of small world networks, which are represented in Figure 3.3. He compares the result of three classical models of opinion dynamics when they are implemented with these networks. The surprising (and rather worrying) conclusion is that it is not possible to generalize the results obtained with one type of network to the other.

Cognitive biases Another endeavor of opinion diffusion models is to study how the characteristics of individuals affect the dynamics of opinions at the group level. Among these individual characteristics, **cognitive biases** are of particular interest to modelers, as they may have important impacts on opinion dynamics. Cognitive biases are defined as systematic features of reasoning which deviate from a rational norm. Classifications of cognitive biases typically list between a few dozens and a hundred of them (Benson and Manoogian, 2017; Ellis, 2018; Arnott, 1998). This high number raises the question of whether there is a problem with our definition of "rational judgement", and new approaches explore the idea that what appears to be individual cognitive biases are actually rational features of judgement which have evolved in specific contexts (Haselton et al., 2015). For instance, cognitive scientists Hugo Mercier and Dan Sperber defend that "myside bias" is a bias which originated from the need to exchange argument in social groups (Mercier and Sperber, 2017).

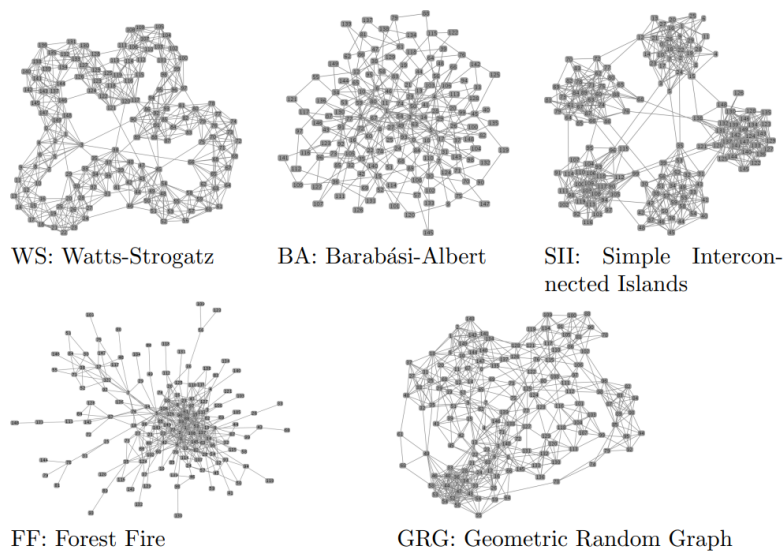


Figure 3.3: The five kinds of small world networks identified by [Thiriot \(2020\)](#)

Confirmation bias is the tendency for people to primarily look for and accept new information which confirms their previously held opinion. Confirmation bias is commonly modeled in ABMs of opinion dynamics ([O'Connor and Weatherall, 2018](#); [Dupuis de Tarlé et al., 2022a](#)), in order to test its effect and as a potential mechanism causing polarization.

Myside bias is defined by [Mercier and Sperber \(2017\)](#) as a tendency to produce reasons in favor of one's point of view. Mercier and Sperber defend the idea that myside bias is the result of an evolution mechanism which favored our capacity to convince our peers. They argue that it causes problems for individual reasoning but that its effects are mitigated in social contexts, where it allows a cognitive division of labour. They cite scientific communities as an example of a social contexts where the effect of the bias is mitigated. In Chapter 7, I present an agent-based model studying the impact of myside bias on scientific communities.

3.2.2 Classical models of opinion

The voter model One of the first agent-based models of opinions is the voter model introduced by [Clifford and Sudbury \(1973\)](#) and [Holley and Liggett \(1975\)](#). In this rather simple ABM, agents are "voters" connected in a graph, whose opinion can take one of two values, 0 or 1. The opinion of the agents change depending on that of their neighbours: at each time step a random agent is selected along with one of its neighbors, and the agent takes the opinion of the neighbor. This model was quite successful because it has an equivalence with physics model, and its differential equations can be solved ([Castellano et al., 2009](#)). Studies of this model have shown that the results highly depend on the structure of the graph connecting the agents. For instance, in a lattice (multi-dimensional grid), consensus is always

reached if the system is finite. Variants of the model have been studied, including introducing one or several "zealot" agents, who never change their opinion: [Mobilia \(2003\)](#) have shown that even a single zealot can influence all the network. [Vazquez et al. \(2003\)](#) have proposed a versions of the model where the agents can have three states of opinion: centrist, left and right, where the extremes only interact with the centrist: this creates a discrete version of the bounded confidence model (see below).

Axelrod model of culture differentiation The model of [Axelrod \(1997b\)](#) can be considered a precursor of opinion diffusion models. This model studies how homogeneous "cultures" can emerge. It is a spatial model: agents are arranged on a grid, and each agent interacts only with its neighbors. The cultural characteristics of agents are features, each of which may be one of multiple possible traits. Agents interact with their direct neighbors and imitate them with a probability that is equal to their similarity: they interact more with those like themselves and tend to become more like those with whom they interact. As a result, large portions of the grid end up being composed of agents who are culturally similar, i.e. who share a great number of traits. This model was followed by significant further contributions in [Flache and Macy \(2007\)](#) and [Centola et al. \(2007\)](#).

Bounded-confidence models The model of [Hegselmann and Krause \(2002\)](#) is a very influential model of belief change and opinion polarization. This model represents homophily by introducing the concept of **bounded confidence**. The model of opinion is very simple: opinions are numbers in the $[0, 1]$ interval, which may be interpreted as a position on a spectrum (e.g. a political spectrum). The way agents update their beliefs is by taking an average of the opinions that are "close enough" to their own. It is as if the agents only trust those whose opinions are not too different: this selected updating is denoted "bounded confidence". Because the opinions are numbers between 0 and 1, it is possible to represent their evolution in opinion profiles such as those presented in [Figure 3.4](#), where the x-axis represents time steps of the simulation and the y-axis the opinion of each agent. The threshold of what counts as "close enough" for influence is a crucial parameter in the model. Notably, with certain values of this threshold, the agents form two distinct opinions. [Figure 3.4](#) presents an example of this phenomenon. Thus, this model exhibits polarization under certain conditions, in the sense of "bimodality" defined by [Bramson et al. \(2017a\)](#).

[Deffuant et al. \(2000, 2002\)](#) created an important extension of the bounded confidence model of Hegselmann and Krause. They introduce *continuous* influence values: instead of a threshold separating between a strong influence and no influence at all, in Deffuant et al. all agents influence each other, proportionally to the overlap of their opinions. Indeed, they add a notion of "uncertainty" range: each agent is equipped with an opinion and an uncertainty interval, both of which are updated at each turn. Agents with a small uncertainty interval are less likely to be influenced, and conversely.

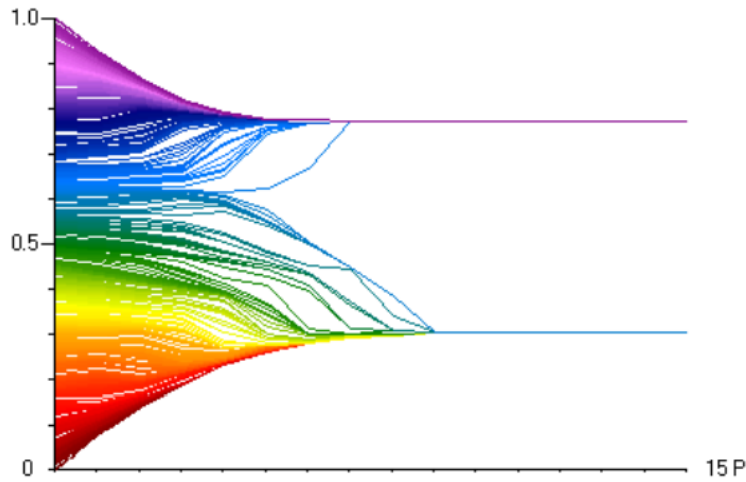


Figure 3.4: Example of the evolution of the opinion of agents in the model of Hegselmann and Krause (Hegselmann and Krause, 2002).

3.2.3 Modelling opinion with arguments

While argumentative exchange has long been neglected in the agent-based modeling community, in recent years it has gained increasing attention from scholars across diverse domains, who seek to enhance opinion dynamics by representing the change of opinions through the exchange of "reasons", or arguments. Unlike classical models, where opinions are influenced solely by social factors, in argumentative models, opinions are shaped by reasoned arguments, and agents engage in reasoning processes to form opinions rather than simply conforming to their social environment. The question is whether we observe different dynamics, or reproduce known dynamics, once agents do not (only) exchange opinions, but also exchange arguments.

An exhaustive review of all the ABMs which use argumentation is beyond the scope of this work. In the following section, I present a few influential models, referring the reader to Dupuis de Tarlé et al. (ming) for a complete overview of the domain. In line with the methodology of Dupuis de Tarlé et al. (ming), I consider opinion dynamics models which include both explicit representation of arguments and computer simulations. I omit the models which use abstract argumentation graphs and semantics to model the cognition of agents here, such as (Taillandier et al., 2021; Butler et al., 2019; Gabbriellini and Torroni, 2014): these models are the closest to my work, and I dedicate Chapter 4 to their in-depth study and comparison.

The Argument Communication Theory of Bipolarization of Mäs and Flache The first authors to explicitly model an argument exchange in an ABM of opinion dynamics are Mäs and Flache (2013). Their goal is to model bipolarization, which is a form of polarization where groups with opposing views become increasingly extreme. Traditional sociological theories could not fully explain bipolarization, which led to the proposal of three potential mechanisms: positive social

influence, homophily, and negative social influence. However, none could account for the radicalization of opinions. To address this gap, Mäs and Flache propose the Argument Communication Theory of Bipolarization (ACTB). Their idea is that when like-minded individuals exchange reasons rather than mere opinions, it leads to the formation of polarized groups with increasingly extreme views. They test this theory through an argumentative agent-based model, which they call the AC model. In the AC model, the agents debate over a single issue i . Each argument either fully supports the issue or its opposite. The opinion of agents are numerically represented as values in $[-1, 1]$ (where -1 is strictly disfavoring i while 1 favors i). In this way, the results of the AC model can easily be compared with those of classical models of opinion dynamics (Hegselmann and Krause, 2002; Deffuant et al., 2000). Another similarity is that the model implements homophily: agents exchange arguments only with those who have a similar opinion. The authors find that the agents do polarize into two groups who progressively become more and more extreme. The AC model has proved quite influential, as many authors have expanded on it, to investigate social networks (Mäs and Bischofberger, 2015), the biased processing of arguments (Liu et al., 2015), forgetting mechanisms (Singer et al., 2019), or real-world debates (Banisch and Shamon, 2022).

Structured arguments Kopecky (2022) investigates polarization effects in the absence of both negative influence and homophily. The underlying question is whether different types of argumentative strategies give rise to polarization or have the potential to de-polarize a society of agents. For this, Kopecky introduces arguments which are more descriptive than those of the AC model or abstract argumentation. The arguments are *structured*: they are formed of several literals, the premises, and one literal which is the conclusion. The conclusion of an argument a can be one of the premises of another argument b , in which case a supports b , or the negation of a premise of b , in which case a attacks b . Thus, the relations between arguments form a bipolar argumentation graph.

The strategies defined by Kopecky are of two types:

- *Allocentric* strategies corresponds to generating arguments whose premises are accepted by the opponent. For instance, according to the *convert strategy* the proponent produces an argument whose premises are accepted by the opponent and whose conclusion is accepted by the proponent.
- *Egocentric* strategies corresponds to the case where agents produce arguments based on premises that they themselves accept, without taking their opponent into consideration.

Using three polarization measures from (Bramson et al., 2017a), Kopecky shows that allocentric strategies have de-polarizing effects, while egocentric strategies tend to polarize. This model shows that one must not need to assume homophily or

confirmation bias to observe polarization, and that argumentative strategies have a non-trivial impact on the opinion dynamics.

Argumentation with natural language An interesting take on ABMs of argumentation is that of [Betz \(2022\)](#), whose model’s arguments are generated by a large language model (LLM), specifically GPT-2 ([Radford et al., 2019](#)). To my knowledge, Betz is the only argumentative ABM which uses natural language. Agents exchange arguments for and against a given issue in the form of posts generated by the LLM. The memory of the agents is initialized with posts from a Kialo ([kialo.com](#)) debate on the legalization of drugs. Their memory allows them to keep track of past posts, but it is limited. The opinion of an agent $\mathbf{ag}$ at time point t on the given issue is calculated on the basis of the conditional probability of producing a claim for (**pro**) resp. against (**con**) the given issue:

$$\text{opinion}(\text{memory}_{\mathbf{ag}}^t) = \frac{1 - P(\text{con} \mid \text{memory}_{\mathbf{ag}}^t)}{(1 - P(\text{pro} \mid \text{memory}_{\mathbf{ag}}^t)) + (1 - P(\text{con} \mid \text{memory}_{\mathbf{ag}}^t))}$$

The communication between agents can take one of three forms: updating on all new posts (baseline scenario), likelihood-based homophily with bounded confidence, or confirmation bias. Agents have different strategies: passive agents share posts from their memory but that do not create new posts, and creative agents generate new (more or less surprising) posts.

Generative large language models can be considered one form of simulation of human cognition. Thus, comparing this model’s results with that of other opinion dynamics models (with or without explicit argumentation) is a step towards validating the robustness of their results. Where there are only passive agents, the model yields dynamics similar to other opinion dynamics models such as ([Mäs and Flache, 2013](#); [Hegselmann and Krause, 2002](#); [Singer et al., 2019](#)): no polarization in the baseline scenario, strongest polarization for homophily. When there are creative agents, however, differences appear: stronger clustering already in the baseline scenario, but less clustering for homophily. This hints that the *production* of arguments, which is rarely represented in argumentative ABMs, may be an important aspect of the opinion dynamics.

3.3 Models of Epistemic Communities

Computational social epistemology is a branch of agent-based modeling which studies the ability of agents to discover the best alternative amongst several possibilities. Alvin Goldman introduces the concept of social epistemology in his 1999 book "Knowledge in a social world" ([Goldman, 1999](#)), where he proposes as an object of study the way in which social processes influence our acquisition of true beliefs. In this book, Goldman sketches a definition of the concept of truth. He introduces the

main competing theories about the nature of truth in epistemology; then criticises them. He defends the idea of a transcendent truth and favors the **correspondence theory of truth**, which states that the truth property reflects a correspondence between what is said and reality.

Goldman sketches out a formal framework for evaluating epistemic practices which constitutes the basis of computational social epistemology. Goldman's framework includes degree of beliefs for the agents which are numbers between 0 and 1, and he uses an objective truth to evaluate individual and social practices.

"Epistemic communities" are communities of agents whose aim is to discover the truth about an issue. Scientific communities are often the typical example of epistemic communities. Contrary to the opinion dynamics models presented so far, in models of epistemic communities all opinions are not considered of equal worth. Models of epistemic communities propose different definitions of what constitutes "the truth", or the best alternative, often following the correspondence theory of truth. Here, we can also adopt a modeler's definition, and define the truth as the best alternative that the agents can choose given the context of the simulation. The epistemic performance of a community of agents can be measured, for instance as the proportion of agents whose position reflect the truth. In the following section, we describe a few influential models of epistemic communities.

3.3.1 Truth in the bounded confidence model

Hegselmann and Krause extended their model in ([Hegselmann et al., 2006](#)) to a situation where one opinion represents a ground truth. This truth value "attracts" the opinion of some agents, the "truth-seekers". Their model contains agents who are truth seekers, and agents who are not, and all agents participate in social interactions together. This allows them to investigate a cognitive division of labour between truth-seekers and non-truth seekers. They show that when a community is composed only of truth seekers, it always converges to the truth. Figure 3.5 represents this situation, where the truth value is the dashed line. They also show that the position of the truth in the opinion space is quite important; it can be rather central (close to 0.5) or extreme (close to 0 or 1). In particular, when there are non-truth seekers and the truth is extreme, it is harder for all agents to converge to the truth: non truth seekers who start with the opposite opinion sometimes get stuck with wrong opinions. Figure 3.6 exemplifies such a situation.

They report several interesting results.

"Leading the pack" theorem The authors show that even one truth-seeking agent can lead the whole community to the truth, as long as the value of the threshold allows all agents to listen (even indirectly) to her.

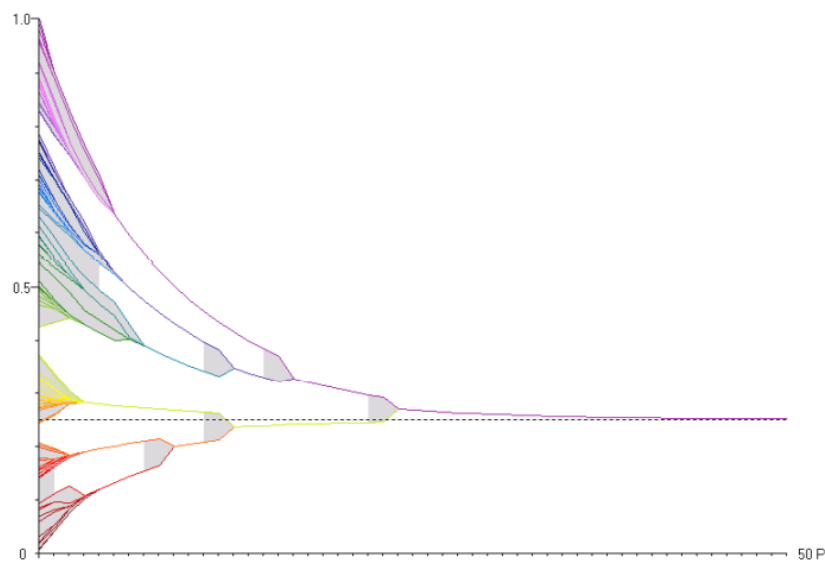


Figure 3.5: Example of the formation of a consensus to the truth value in the updated model of Hegselmann et al. (2006). The opinion of the agents are the full lines, and the dashed line represents the truth value.

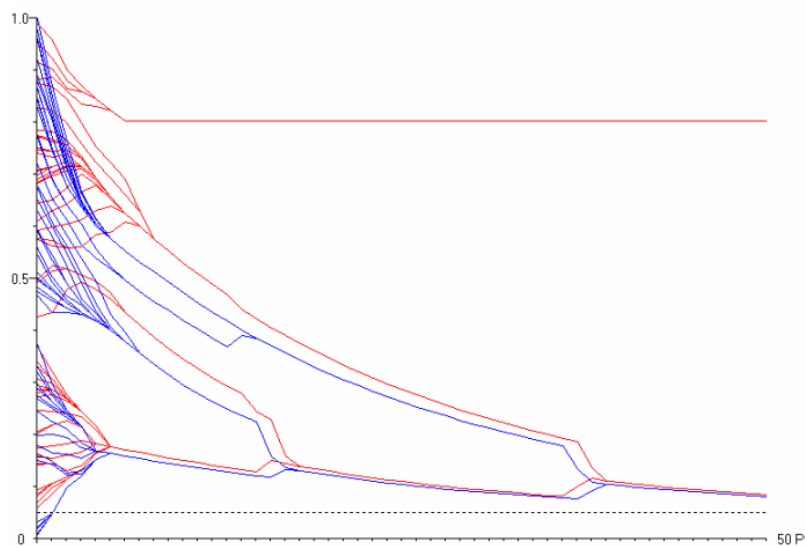


Figure 3.6: Example of an opinion profile in the model of Hegselmann et al. (2006). The opinion of the agents are the full lines, with truth seekers in blue and non truth seekers in red, and the dashed line represents the truth value.

"Leave behind" effect Figure 3.7 represent what they analyse as the **leaving behind effect**. One could think intuitively that the higher the attraction to the truth, the higher the chance of consensus to the truth. This is not the case in this graph. It is because of the bounded confidence effect and the fact that only some agents are truth seekers: a small group of truth seekers could move to an extreme truth too fast, thus leaving behind the rest of the population who would stop listening to them. This is quite interesting: there are many situation in the history

of science where a small group which held the truth failed to convince public opinion. One could think of the first scientists to identify global warming for instance.

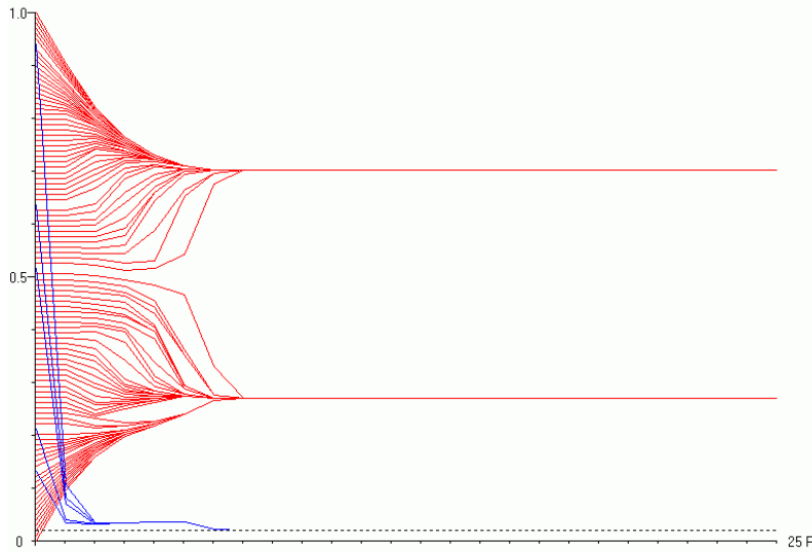


Figure 3.7: Example of an opinion profile in the model of [Hegselmann et al. \(2006\)](#) which exemplifies the "leave behind" effect. The opinion of the agents are the full lines, with truth seekers in blue and non truth seekers in red, and the dashed line represents the truth value.

3.3.2 Bayesian networks

Zollman model [Zollman \(2007\)](#) adapts a model created by the economists [Bala and Goyal \(1998\)](#) to represent scientific communities. He aims to study how the way information is transferred impacts the epistemic performance of a community. Zollman's model is based on the assumption that scientists are not merely passive observers of the truth, but that they experiment with their environment to produce evidence. He models two competing alternatives, one of which is preferable to the other, using "one-armed bandits": each alternative is characterized by a payoff value, which represents the average reward an agent would receive by testing it. Formally, the alternatives are modeled as Gaussian distribution centered on the mean payoff value. For instance, if an agent tests option A with a payoff of 0.5 five times in a row, she may receive the values (0.51, 0.45, 0.55, 0.52, 0.49). The agents form a preference between the two alternatives by using Bayesian reasoning; thus, these models are often called "Bayesian networks". They start with a random belief about the payoff value of each alternative, and update it when they obtain new evidence, either by testing it or by receiving it from other agents. Agents succeed when they all choose the best alternative.

Agents are part of a social network and share all their evidence with their neighbors. Zollman studies the impact of *connectivity* (the number of connections between agents) on the epistemic performance of the community. For this, he tests his model on three types of networks: the cycle, the wheel and the complete graph,

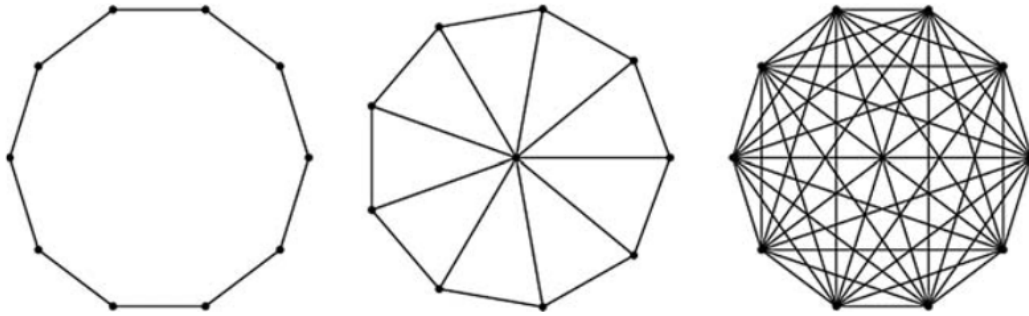


Figure 3.8: Networks used in [Zollman \(2007\)](#) illustrated for 10 agents: from left to right, the cycle, the wheel and the complete graph.

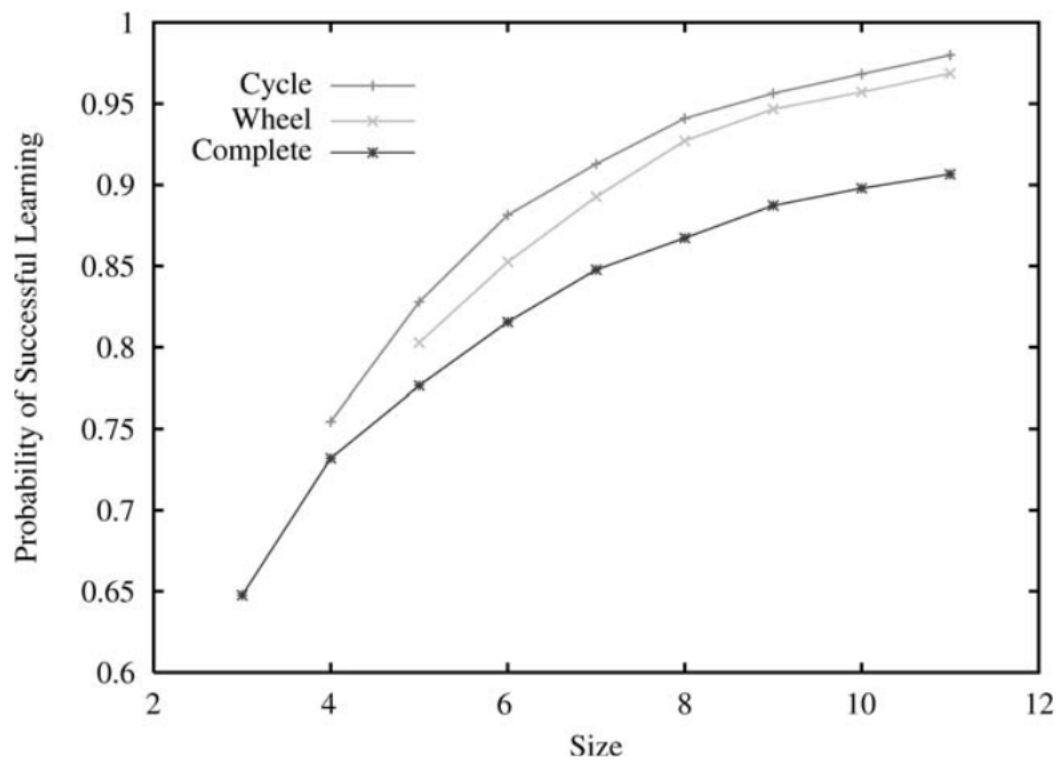


Figure 3.9: Results obtained in the model of [Zollman \(2007\)](#): the frequency of success of the community as a function of its size and social network.

as represented on Figure 3.8. The results, shown in Figure 3.9, are remarkably counter-intuitive: the wheel performs better than the cycle, which performs better than the complete graph, for all size tested. This is what has been called the **Zollman effect**: less connectivity between agents implies better (though slower) outcomes. This effect is due to the fact that limiting the communications between agents preserves a certain form of *diversity*. Agents sometimes do not succeed because they all abandon the best alternative after a few misleading results. If they do not communicate much, there is a higher chance that some individuals will continue working on the best theory. Zollman analyses this result in relation to Kitcher’s “division of cognitive labor” (Kitcher, 1990), which states that scientists have to work on different theories in order to optimize the chance that the community finds the optimal response, even though some of them will work on bad alternatives. In Zollman’s model, the lack of connectivity creates diversity which allows for a natural division of labor, as scientists pursue the alternative which they find most promising. Rosenstock et al. (2017) have shown that the effect of the connectivity of a network (the Zollman effect) are **not** robust, in particular to size effects.

Scientific polarization Zollman networks have been used by philosopher Cailin O’Connor and colleagues to investigate features of opinion dynamics in scientific communities.

O’Connor and Weatherall (2018) use Bayesian networks to model scientific communities where polarization occurs even in the face of actual evidence. For this, they model a lack of trust between agents. The agents update their beliefs with a function which treats evidence from others as uncertain depending on the absolute distance between their credence in a certain theory. Their modelling is inspired by Jeffrey’s conditionalization, which extends Bayesian conditionalization to uncertain evidence. Two mechanisms are explored:

- Anti-updating, when agents receive evidence from agents whose credences are very different from their own, this will lessen their initial credence in this evidence. Thus this evidence can have the effect of strengthening their own credence.
- No anti-updating : Agents simply ignore all evidence from agents whose credences are sufficiently different from their own. The second mechanism effectively behaves like a bounded confidence after a certain threshold, whose size is determined by function parameter m . However, the overall mechanism is more complex, as even when agents are still within acceptable distance, they influence each other less as the distance increases.

The authors use a complete network to model the social network of the agents. This means that all agents interact with all others and polarization occurs because of a lack of trust and not a differential access to information. All of the agent’s opinions

are assigned randomly at the beginning. They find that mistrusting others based on their beliefs leaves the community in a worse epistemic state. They temper this statement by pointing out that the agents here are all in good epistemic conditions as opposed to a situation where it might be more beneficial to distrust, for instance in the case of biased agents. The anti-updating mechanism leads to more polarization.

Propaganda and manipulation using science In (Weatherall et al., 2020), Weatherall, O'Connor and Bruner present a “tobacco strategy” model. Described by Oreskes and Conway in their book “Merchants of Doubts” (Oreskes and Conway, 2011), the tobacco strategy was developed by the tobacco industry to influence political decision-makers and prevent the introduction of regulations due to the health risks of tobacco. This strategy, which Oreskes and Conway also associate with climate deniers, consists of selecting, sharing and funding research that supports the industry’s privileged position - for example, that tobacco is not harmful.

Weatherall et al. (2020) model this strategy within Bayesian networks similar to Zollman’s. For this, they add two new kinds of agents: *policy makers* are agents which update their beliefs upon the evidence that they receive (e.g. from scientists) but do not gather evidence themselves, and *propagandists* are agents which receive evidence from all the scientists and selects which one they transfer to the policy-makers. Figure 3.10 represents the network which links the three kinds of agents: policy makers, scientists and propagandist.

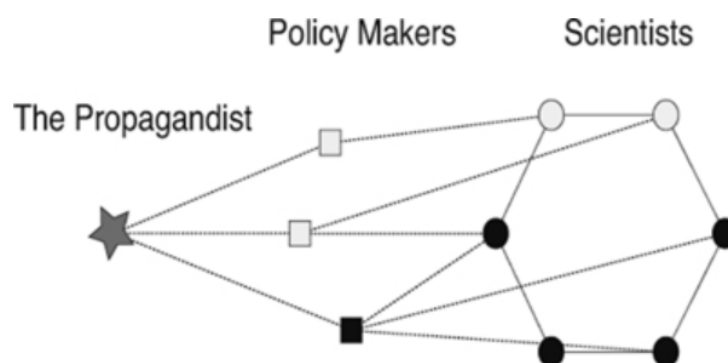


Figure 3.10: Network which links the three kinds of agents: policy makers, scientists and propagandist, in the model of (Weatherall et al., 2020).

They model a first version of the strategy, **selective sharing**, by having the propagandist share only results which supports one theory over the other. Thus, the policymakers receive results which are biased in favor of one option, even though all scientists are unbiased and produce “good” evidence. This strategy relies on the fact that in science, there are some statistical effects which means that some results will not support the “true” theory: not everyone who smokes will have a cancer and thus there may be (a few) studies which will not show a link between cigarettes and cancer. This is why meta-studies are needed to evaluate a theory based on all

available evidence. In the model, this stochasticity is represented by the one-armed bandits whose results are only supporting the true belief **on average**.

A second strategy, denoted **biased production**, corresponds to industrial actors funding their own science. It is modelled by allowing the propagandist agent to gather her own evidence, and then share only the results which favor the wrong theory. Note that in this case, the propagandist is not doing "bad science" e.g. intentionally manipulating the results, she is only selectively choosing not to publish some results.

Finally, the authors model a situation where journalists can unwittingly play the role of propagandists. In the US, between 1949 to 1987, medias had to obey the "Fairness Doctrine", which required them to give equal time to multiple perspective on an issue. This doctrine contributes to the creation of an appearance of scientific controversy even when there is a scientific consensus, as was shown in the case of climate change and tobacco by [Oreskes and Conway \(2011\)](#). It can be used to manipulate public opinion by presenting the two sides of a question as equally strong. In the model, *journalist* agents act as a filter and transfer an equal number of evidence in favor of each theory to policymakers.

[Weatherall et al. \(2020\)](#) show that each of these three situations leads to the manipulation of the opinion of policymakers in favor of the false belief, even though there is a consensus of the scientists on the true belief.

Modelling Myside Bias [Baccini et al. \(2023\)](#) explore the effect of the combination of a production bias and an evaluative one in a context of deliberation among Bayesian agents. They develop and justify a Bayesian model of myside bias at the level of individual reasoning ([Baccini and Hartmann, 2022](#)), and then study how this may affect group deliberation.

First, they represent biased agents through the Bayesian notion of diagnosticity ([Hahn and Hornikx, 2016](#)). The diagnosticity of a proposition A with respect to a proposition B corresponds to how indicative is proposition A of whether or not proposition B is true. This is formalized by the likelihood ratio of A with respect to B . [Baccini et al. \(2023\)](#) represent biased agents as agents who perceive the likelihood of a certain argument differently based on whether or not such argument is in favour or against their initial hypothesis. They attribute lower diagnosticity to arguments that are against their present position, and higher diagnosticity to arguments in favour. This captures the notion of myside bias as an evaluative bias ([Stanovich et al., 2013](#)). In addition, they also allow agents to only share with other agents arguments that are in favour of their position.

Based on this, they study whether or not a community is able to still determine whether a certain hypothesis is true or not. They assume that agents are already competent on the matter, that is, that agents assign to the correct hypothesis at

stake a probability on average higher than half. Thus, they mainly observe two possible outcomes, that is bi-polarization and consensus. In particular, they find that when the radicality of bias is distributed equally among agents, correct consensus is more likely when the community is smaller, and when the strength of the bias is weaker. Small groups of biased agents can be better at collectively tracking the truth than non-biased agents. This is because the bias makes such communities less prone to be swayed by misleading arguments, and consequently preserves their own initial competence. However, the picture changes radically when the biased agents are those starting with low levels of competence. In that case, the presence of bias makes communities way less effective in tracking truth.

3.3.3 ArgABM: modelling scientific inquiry with argumentation

The argumentative agent based model **ArgABM** of Borg et al. ([Borg et al., 2017a,b, 2018, 2019](#)) is motivated by the idea that scientific interaction among proponents of rival scientific theories is argumentative in nature. Scientists do not exchange evidence, like they did in the Bayesian models, but rather arguments for and against their theories. In light of the available arguments, they determine which of the theories is better in the sense of being more ‘defensible’. In the context of scientific inquiry, arguments are akin to studies supporting a theory, which can be attacked by studies from rival theories but defended by further studies within the same theory. The model has been used to study how different aspects of individual inquiry impact the efficiency of knowledge acquisition. More precisely, Borg and colleagues investigate how:

1. different social networks among scientists (with varying degrees of connectivity),
2. different types of information sharing (such as reliable and deceptive),
3. different norms of inquiry (such as cautious or incautious decision-making about changing one’s current theory), and
4. different norms for evaluating scientific theories (such as measuring the degree to which a theory is ‘defensible’, or the degree to which it is ‘anomalous’)

impact the efficiency of the community in reaching consensus on the best of the available theories.

ArgABM utilizes an "argumentative landscape" to simulate argumentative exchanges, where rival theories in a domain are represented as rooted trees. Arguments within these trees are connected by a "discovery relation", depicting the path agents

take from one argument to another. Arguments from one theory can attack arguments from another. This representation is based on Dung’s abstract argumentation, enriched with the discovery relation. Scientists explore this landscape, discovering arguments and attacks, aiming to determine the most defensible theory. There is only one theory which can withstands all attacks: agents are successful when they converge on that theory.

As mentioned above, the goal of simulations is to assess how different socio-epistemic factors impact the success of inquiry. At the start of a simulation scientists are randomly positioned at the roots of theories. From thereon, they (i) explore the landscape, (ii) exchange information within their social network and (iii) update their assessment of the given theories. Finally, every few rounds scientists evaluate the theories to decide which one they should pursue. The simulations show that the more connected a scientific community is, the more likely it is to converge on the best theory, under a variety of conditions.

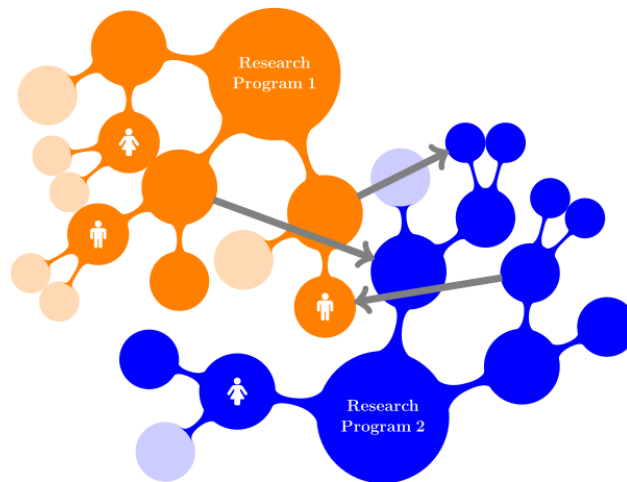


Figure 3.11: An example of an argumentative landscape consisting of 2 theories (or research programs). Darker shaded nodes represent arguments that have been investigated by agents and are thus visible to them; brighter shaded nodes stand for arguments that are not visible to agents. The largest node in each theory is the root argument, from which agents start their exploration via the discovery relation, which connects arguments within one theory. Arrows represent attacks from an argument in one theory to an argument in another theory (Borg et al., 2019).

Agent-Based Models of Abstract Argumentation

Argumentation is a form of reasoning that is inherently *social*. It can serve various purposes, from persuading others of a particular viewpoint to collective decision-making and problem-solving. Given this collaborative dimension, agent-based models (ABMs) are particularly well-suited for exploring research questions related to argumentation. In the following, I refer to "Argumentative ABMs" as agent-based models that simulate the exchange of arguments between agents. In these models, agents are equipped with mechanisms for exchanging arguments, often using the formalism of **abstract argumentation** (see Chapter 2).

Argumentative ABMs can serve several distinct purposes. On the one hand, these models can help us understand the dynamics of social situations which are explicitly argumentative. This would be the case of models whose goal is to shed light on situations such as political deliberation, collective decision making, or scientific controversies: for instance, [Borg et al. \(2017a\)](#) used an agent-based model to study debates between scientists. Argumentation may also be used as a tool to enhance opinion diffusion models: indeed, it provides a more realistic account of how people change their minds than what is typically assumed in the "classical" opinion diffusion models, and allows for more complex dynamics to emerge. For instance, [Taillandier et al. \(2021\)](#) motivate their work as an extension of previous models of opinion dynamics. Another way that argumentative ABMs can be used is to help us understand the tools of formal argumentation themselves: for instance, in [Dupuis de Tarlé et al. \(2022a\)](#) we used an ABM to study how a specific tool, a gradual semantic, behaved in a collective setting and what kind of dynamics emerged. Note that these goals are not mutually exclusive and that some models explore several of these aspects: for example, [Butler et al. \(2019\)](#) propose a model which represents collective decision-making, attempting to breach a gap between deliberation (an explicitly argumentative situation) and opinion diffusion.

There has been a recent increase in the number of argumentative ABMs. I refer the reader to [Dupuis de Tarlé et al. \(forthcoming\)](#) for a complete overview of this literature. In particular, many models are inspired by the influential model of [Mäs et al. \(2013\)](#) (see Chapter 3). In this chapter however, I chose to narrow my focus and describe exclusively ABMs which *explicitly* state that they use abstract argumentation, where the arguments are modelled using Dung argumentation graphs and the reasoning is modelled with semantics. I made this choice because these models were more closely related and influential in the construction of my own, and because it allowed me to make more fruitful comparison between the diverse modelling choices.

Paper	Year	Research Question
Gabbriellini and Torroni	2013	Connect two streams of research in social simulation: Propose a model which represents the cognitive capabilities of socially embedded agents.
Borg et al.	2017	Model scientific inquiry through the exchange of arguments between scientists, to investigate the role of socio-epistemic factors.
Butler et al.	2019	Breach a gap between deliberation and opinion diffusion by modeling both processes.
Taillandier et al.	2021	Propose an enhanced opinion diffusion model using argumentation, and apply it to the diffusion of vegetarian diets.
Ledoyen et al.	2021	Adapt the model of Taillandier et al. to represent innovation diffusion.
Dupuis de Tarlé et al.	2022	Understand the dynamics induced by gradual argumentation semantics in social settings.
Rago et al.	2023	Model explanations between human and machines as a conflict resolution using abstract argumentation.

Table 4.1: Agent based models which use abstract argumentation to represent the opinion of agents.

There are models that specify multi-agent protocols using abstract argumentation, but which do not contain simulations: they are analyzed using formal mathematical tools only. For instance, [Marzagão et al. \(2018\)](#), who study a political debate situation, or [Bonzon and Maudet \(2012\)](#), which was influential in the construction of [Dupuis de Tarlé et al. \(2022a\)](#). In both of these models, agents cannot change their position during the debate. This makes a purely formal investigation possible, as the modification of each agent’s opinion graph at each time step would make the study of an entire protocol virtually impossible with formal tools. These models therefore fall outside the scope of the study of opinion dynamics: rather, they study situations of *persuasion*. As such, and in order to be consistent with the definition of agent-based models, which specifically includes simulations, I excluded these models from the following analysis.

In this chapter, I present the fundamental structures which characterize ABMs of abstract argumentation and I develop how they are implemented in the different models of the literature. Section 4.1 exposes how the knowledge of agents is represented and how agents form their point of view on a debate, Section 4.2 delves into

the social structures of these models and how agents communicate with each other. Finally, Section 4.3 presents the results obtained by these argumentative ABMs.

4.1 Knowledge Representation

4.1.1 Mental argumentation frameworks

Abstract argumentation is a useful tool to model agents who reason and exchange arguments. Indeed, graphs are mathematical structures which are easy to implement in most programming languages and agent-based model frameworks, and the semantics of abstract argumentation are well suited for implementation and provide a rich set of tools to build cognitive agents. As such, an increasing number of models use abstract argumentation graphs to represent the internal structure of agents' cognition as well as the way they interact and exchange information. In this section, I present how the models of Table 4.1 use the flexible structure of AFs.

Agents in ABMs are modeled as autonomous and dynamic. This means that each agent has its own knowledge representational module which deals with incoming information and it is used as the basis of agent communication. For argumentation-based ABMs this means that each agent has its own *mental argumentation graph* representing its unique mental model of the debate at a certain time point. The AFs underlying these models feature an explicit attack relation, sometimes coupled with a support relation, and the selection of arguments is governed by an argumentation semantics (see Chapter 2). Such mental argumentation graphs are denoted the "mental maps" of the agents. Mental maps may be understood as the agent's "knowledge base", the arguments and relations that they consider to form a position regarding the topic of the debate. The mental AFs are specific to each agent since they may have different insights, and relative to time points since agents may learn from each other or from their environment, as well as forget arguments. In each round an agent may be exposed to new information (e.g., based on communication with other agents or interaction with its environment) and may therefore update its mental AF by incorporating new attacks and/or new relations.

Some models define a **global argumentation framework** which represents all the arguments and attack/support relations which form a debate. In these models, agents are equipped with a **subset** of this argumentation framework, and they all agree on the relations between the arguments. Borg et al. (2017a), Butler et al. (2019), Dupuis de Tarlé et al. (2022a) and Rago et al. (2023) all follow this principle. In these models, the global graphs are randomly generated at the beginning of every simulation.

The model of Dupuis de Tarlé et al. (2022a) implements this principle, using attack-only argumentation frameworks. The global graph of their debates, and

every agent's mental map which are subsets of this global graph have a specific structure: they are "issue-oriented argumentation frameworks" (IOAF). An IOAF is a graph where a special argument, the *issue*, forms the root of these graphs, and all arguments are part of a path of attack relations towards the issue. These graphs are also devoid of cycles, which means that they are trees whose in-root is the issue. The issue is the central argument and the subject of the debate: for instance, in a debate about renewable energies, the issue could be "We should use renewable energies.". Figure 4.1 represents the mental map of an agent in this model.

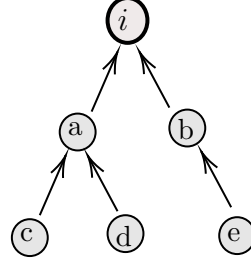


Figure 4.1: Example of a mental argumentation framework of an agent in Dupuis de Tarlé et al. (2022a). Argument i is the *issue* of the debate.

Rago et al. (2023) extend the model of Dupuis de Tarlé et al. (2022a) to simulate the interaction between a human and a machine. Their graphs are also issue-oriented argumentation frameworks, where the issue is called an "explanandum". Their AFs are bipolar argumentation frameworks (and not attack-only like the other models of Table 4.1), and they can be equipped with weights (denoted "base scores" in the article). Weights are numbers in a given interval $\mathbb{I}$ (in their simulations, $\mathbb{I} = [0, 1]$) which represent an intrinsic quality of an argument. They model the notion that not all arguments are to be evaluated similarly. Figure 4.2 represents the mental map of an agent in Rago et al. (2023), where full arrows represent attack relations, and dashed arrows represent support relations between arguments.

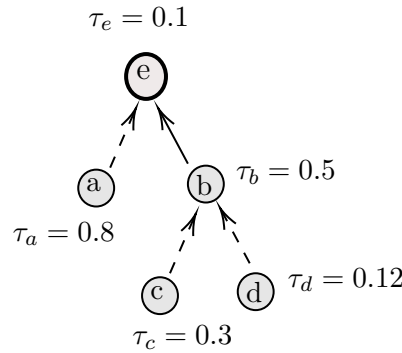


Figure 4.2: Example of a mental argumentation framework of an agent in Rago et al. (2023). Argument e is the *explanandum*. Every argument a is equipped with a base score τ_a . Full arrow indicate attack relation, and dashed arrows indicate support relations between arguments.

Borg et al. (2017a) is a peculiar instance of this principle. The model is characterized by a global "argumentative landscape": a fixed argumentation graph which contains arguments and attacks between them, equipped with a specific structure in

order to represent the process of scientific inquiry. In addition to the attack relation, some arguments are linked together by a "discovery" relation: if arguments a and b are linked by this relation, it means that an agent can "discover" argument b by exploring argument a . The global graph is composed of theories, which are sets of conflict-free arguments which are linked together by discovery relations. Although they do not explicitly describe them as mental maps of the agents, the authors specify that "an agent operates on her own (subjective) fragment of the landscape, which consists of her discovered arguments which are explored by her to a specific degree, and her discovered (attack and discovery) relations between the arguments" (Borg et al., 2017a). Thus, the arguments known by the agents are characterized by a degree of exploration which specifies how much the agent has explored the relations leading to or from the argument. Figure 4.3 represents an example of a mental argumentation framework in the model of Borg et al. (2017a).

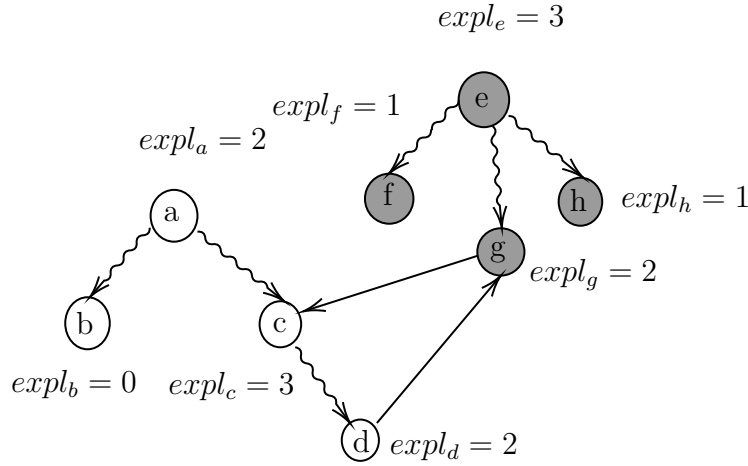


Figure 4.3: Example of the mental representation of the debate for an agent of Borg et al. (2017a). White and grey sets of arguments respectively represent competing theories. Arguments from distinct theory can attack each other, represented by straight arrow. Wavy arrows represent the discovery relation which links arguments from the same theory. Each argument is characterized by an exploration score.

In Butler et al. (2019), the agents interact within two modes: they influence each other's opinion in pairwise interaction phases, and exchange arguments in deliberation phases. During the deliberation phase, the agents are equipped with a set of arguments and attack relations which form their mental argumentation framework. The arguments are characterized by a "degree of adherence" $v \in [-1, 1]$ to an abstract principle $\mathbb{P}$. When v is lower than 0, the argument is against the abstract principle, when it is higher than 0, it is in favor of the abstract principle, and neutral in the case $v = 0$. This abstract principle is the subject of the debate: for instance, $\mathbb{P}$ may be "Liberalism". Figure 4.4 represents the mental map of one of (Butler et al., 2019)'s agents, where the degree of adherence of each argument to the abstract principle is specified.

Gabbriellini and Torroni (2014) depart from these three models in a specific way: in their model, all the agents are aware of the same set of arguments, but they disagree on the attack relations between these arguments. More precisely, each

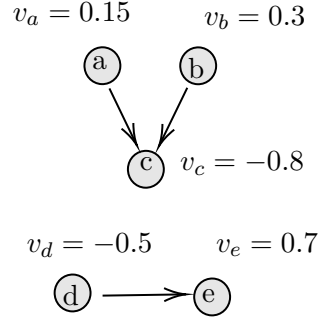


Figure 4.4: Example of the mental argumentation framework of an agent in [Butler et al. \(2019\)](#). Every argument a is equipped with an adherence to abstract principle $\mathbb{P}$ denoted v_a .

agent can be equipped with either one of two sets of attack relations on the same set of arguments. Another specificity is that they do not generate random argument graphs for each simulated debate but rather use the same two possibilities for all their simulations. Figure 4.5 represents the two alternative mental map which are used in their simulations. These graphs were constructed from an online discussion about renewable energies.

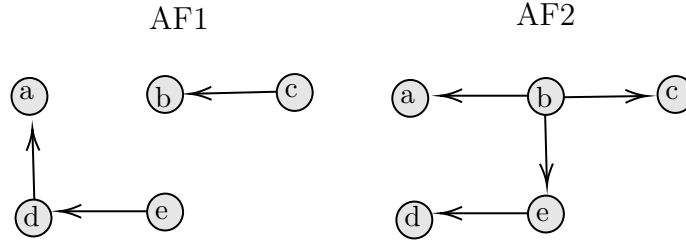


Figure 4.5: Example of the two possible mental argumentation frameworks AF1 and AF2, for an agent in Gabbriellini and Torroni ([Gabbriellini and Torroni, 2014](#)).

[Taillandier et al. \(2021\)](#) present a model whose aim is to study the diffusion of vegetarian diet within a population. Their agent's mental frameworks are attack-only AFs with several specifics. First, they equip every argument with many informations: the option (diet) concerned by the argument and the type of the argument (whether it is pro, con or neutral towards the option), the type of source that the argument comes from and the agents who proposed it, the statement (i.e. conclusion) and rationale (i.e. hypothesis) of the argument, and the importance of each criterion (e.g., "Nutritional", "Economic", "Environmental") on which the argument relies on. For instance, an argument in favor of veganism may rely on a environmentalist value, while an argument in favor of meat consumption may rely on a nutritional value. They provide this level of detail because they built an argument database by extracting arguments about vegetarianism on the internet. In their simulations, the arguments are less detailed: all of the arguments concern only one option, and they are linked to one criterion each. As far as I can tell, the source, statement, and rationale of the arguments do not play any role in the simulations.

Moreover, each criteria has more or less importance to a given agent: an agent may consider the criteria "Nutritional" with a certain importance, and an argument

may rely on this criteria with a certain value. This means that the subjective strength of an argument for an agent ag can be calculated by:

$$\text{str}(ag, a) = \sum_{v \in \text{Values}} \text{importance}(ag, a) \times v(a)$$

They follow the global graph/subset graph scheme, with the exception that agents do not always agree on the relations between arguments. In their global graph, there can be cycles: in particular, two arguments can be linked by a symmetrical attack relation: a attacks b and conversely. However, the mental AFs of the agents do not have cycles: the agents compute a *simplified* subset graph by only retaining the attack coming from the argument with the greatest strength. Thus, agents can both consider distinct arguments and different relations between the same arguments. Figure 4.6 presents an example of the mental map of an agent in Taillandier et al. (2021).

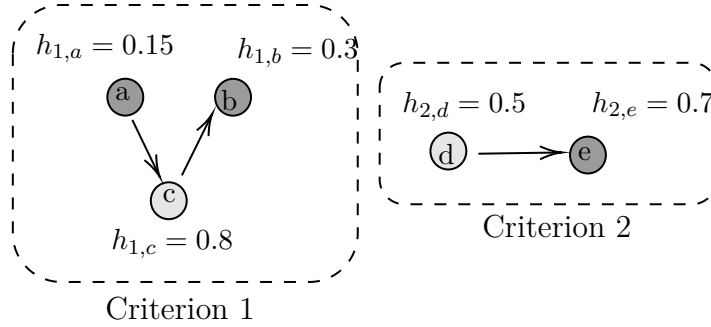


Figure 4.6: Example of the mental map of an agent in Taillandier et al. (2021). Every argument concerns a specific criterion. For every argument a , $h_{i,a}$ represents the importance of criterion i for argument a . For readability, the argument’s strength are not represented on this diagram.

Ledoyen et al. (2021) use the same representation as Taillandier et al., only adding a notion of *trust* in the source of an argument.

When comparing the way the seven models instantiate the mental maps of the agents, we can see that each model implements abstract AF in a different way, whether it is by imposing a structure, or by adding information about arguments such as scores or degrees of adherence. Indeed, many models equip arguments with a number, but this number may reflect very different information: the degree of exploration of arguments of Borg et al. (2017a) has nothing to do with the degree of adherence in Butler et al. (2019). This highlights the versatility of abstract argumentation frameworks, which can be adapted to the modelling of distinct social situations. We can also notice many similarities: for instance the use of values or principles in Butler et al. (2019) and Taillandier et al. (2021): the importance of an argument to a criterion in Taillandier et al. (2021) can be likened to the degree of adherence to a principle which characterize the arguments in Butler et al. (2019). Another noteworthy similarity is that Gabbriellini and Torroni (2014) and Taillandier et al. (2021) are the only models where the AFs are initialized from real data: the argument structure is extracted from a debate about renewable energies

for the former and about vegetarian diets for the latter. These works are also the only one where the agents can disagree on the attack relation between arguments: this may be a hint that real argumentative discussion exhibit this property.

4.1.2 Forming positions on a debate

In argumentative ABMs, the agents are characterised by a *position*: this position can be an opinion, a behavior, a set of arguments, which define how the agent stands regarding the debate. An important feature of the ABMs presented in this chapter is that the position of the agent is derived from the agent's mental map. Like the mental AFs, it evolves throughout the course of the debate as agents learn or forget arguments.

In the models presented in this chapter, the positions of the agents are derived from their personal argumentation graphs using a semantics and, possibly, other additional types of rules. Figure 4.7 illustrates this general principle. As stated in the previous chapter, an abstract argumentation semantics is a function which, applied to an abstract argumentation graph, returns either a set of acceptable argument (in the case of a set-based semantics), or a score of acceptability for each argument (in the case of a gradual semantics). Thus, the position of the agents can take different forms: in some models, it is a numerical opinion, e.g. a number between 0 and 1, and in others, it is a set of arguments, typically an extension returned by a semantics. It can also be a behavior: "vegetarian" or "vegan" for instance, like in the model of Taillandier et al. (2021). Indeed, I define the position as the output of the model which is analysed in the results. In some models, the semantics may return a set of arguments and/or a numerical opinion, but the position may be a combination of these objects and others computed with an adequate set of rules. Models in which the position of the agents is represented by a number have the advantage of a double representation, combining the complexity of the argumentation graph and the simplicity of the one dimensional numeric opinion. These models' output can more easily be compared with that of classical bounded-confidence models.

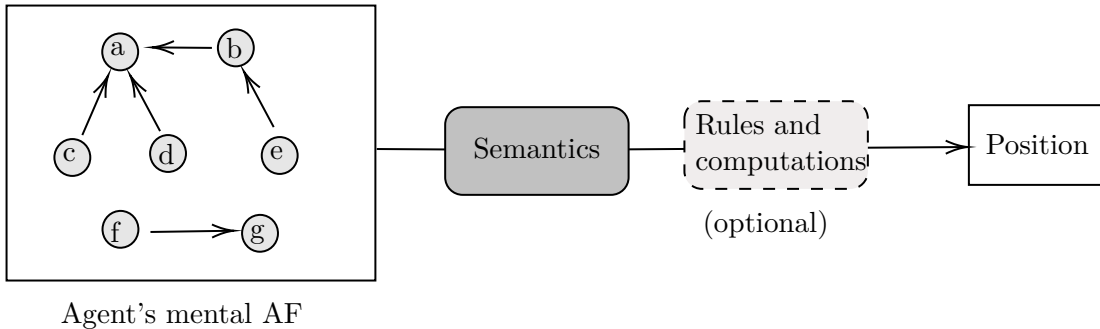


Figure 4.7: Diagram representing how positions are obtained from mental AFs

In Gabbriellini and Torroni (2014), the position of the agent is represented by a set of arguments: the extension returned by the complete semantics. Referring to

the two possible argumentation frameworks represented in Figure 4.5, they assert: "At the beginning, each agent believes either $\{a, c, e\}$ or $\{b, d\}$, the two possible extensions that a complete semantics returns from the AFs used." (Gabbriellini and Torroni, 2014). In the general case, the complete semantics returns multiple extensions, but for graphs with no cycle, like the AFs of Figure 4.5, there exists exactly one extension which is complete, grounded and preferred (Dung, 1995). Thus, the authors track which extension corresponds to the agent's position throughout the debate.

In Borg et al. (2017a), agents' positions are also sets of arguments, although they use a rather original mechanism to specify them. An agent's position is its preferred theory, which is a set of conflict-free argument. An agent investigates arguments from its preferred theory, and can choose to switch its preference to a more "defensible" theory. For this, they compare the *degree of defensibility* of each theory, which corresponds to the number of defended arguments in the theory. As such, this model does not use a semantics, strictly speaking, but a numerical value which does not indicate whether the set of argument itself is acceptable or not. However, as there exists only one theory which completely defends itself, the highest defensibility score will eventually match the notion of acceptability given by classical set-based semantics.

In contrast, in Taillandier et al. (2021), positions are multi-dimensionnal: a numerical opinion, and a discrete behavior, and obtained through several steps. First, the agents determine the preferred extensions from their mental AF. Then, they compute a value for each of these extensions; this value is the (normalized) sum of the strength of arguments favoring a given issue. The extension with the highest **absolute** value is chosen. The value of the chosen extension, a number between -1 and 1 , represents the agents opinion, which is then mapped to a dietary behavior: either "omnivorous", "flexitarian", "vegetarian" or "vegan". This behavior can be compared to their empirical data about the prevalence of various diets. I noted that the fact that the agents choose the preferred extension with the highest absolute value is not discussed, although it seems that it would favor extreme positions. This may be because this is only relevant when the graphs feature cycles: otherwise, the preferred semantics returns only one extension.

Ledoyen et al. (2021) use the same process as Taillandier et al. (2021) for computing the chosen extension and its value. They extend the model by including an "uncertainty". Then, they compute the opinion as an average between the value, denoted "personal benefit" and a perceived "social value". The social value represents what the agent perceives as the opinion of others regarding the innovation. It does not change during the simulation but is initialized for each agent at the beginning, using a gaussian distribution whose parameters are informed by empirical data. Then, the opinion and uncertainty are used to determine the agent's "state of interest" in the innovation. Finally, they use a complex decision process to determine whether the agent adopts the innovation. I refer the reader to the paper for a full detail of the rules applied here.

Dupuis de Tarlé et al. (2022a) define the opinion as a number, just like many classical opinion dynamics models. For this, we apply a gradual semantics to the "issue" of the debate. We use the h-categorizer semantics defined by (Besnard and Hunter, 2001). We can note that the gradual semantics are quite useful, as they allow one to obtain a numerical opinion with all the advantages of its representation without the need to add criteria, importance, strength, etc. Rago et al. (2023) similarly apply a gradual semantics to the root of their argumentation graphs, the "explanandum". Because they use weighted bipolar graphs, they use the semantics Df-QuAD defined by Rago et al. (2016). However, unlike Dupuis de Tarlé et al. (2022a), they do not define their agent's position as continuous opinions but rather "stances" on the explananda. The stance is given by the evaluation of the explanandum by the gradual semantic: if it is strictly higher than 0.5, the stance is positive, if it is strictly lower than 0.5, the stance is negative, and neutral otherwise. Using this, they study the resolution of conflicts between agents: a conflict between two agents (typically, a human and a machine) is resolved if they hold the same stance.

Butler et al. (2019)'s position is a numerical opinion between -1 and 1 . They somewhat depart from the general scheme of Figure 4.7, because the opinion of agents is modified through an alternance of deliberation phases in which agents exchange arguments and direct opinion influence similar to bounded confidence models. Furthermore, the relation between arguments and opinions is more complex: it is the opinion of the agent which determines which arguments she is aware of, and in turn, the result of the collective deliberation influences the agent's opinion.

- **Opinion influence on arguments.** At the beginning of each deliberation phase, each agent receives arguments which reflect her position, with the following principle: there is a certain proportion of arguments whose adherence to the principle is the same as the agent's opinion.
- **Arguments influence on opinions.** In the deliberation phase, agents use the grounded labelling-based semantics to determine the acceptability of arguments, and there is a voting mechanism which leads to arguments being accepted or rejected by the agents. At the end of the deliberation phase, agents update their opinions with the result of the collective deliberation.

4.2 Social Aspects

These models seek to understand the evolution of opinions in the context of social interactions, and as such they include explicit modeling of the social interactions between agents. The rules of social interaction determine how agents give and receive new arguments from the other agents. This is how they update their mental argumentation frameworks and modify their positions, as was detailed in the previous section.

The models that I discuss in this section make distinct choices when it comes to modelling the way agents interact together. The types of interaction between agents can take one of three distinct forms: agents can interact with fixed partners in **fixed social network**, they can meet two-by-two in **pairwise encounters** (also called dyadic interactions), or they can broadcast their arguments directly through **public diffusion**.

Pairwise encounters are often used in the classical opinion diffusion models, such as the bounded confidence models (see Chapter 3). They represent chance encounters between agents which lead to some form of discussion. They consist in meetings between pairs of agents, who change at each turn. The pairs are determined stochastically. They can be completely random, or be determined by some characteristics of the agents, typically the proximity of their opinions.

Social networks are another common way to model social interactions between agents. Social networks represent social links that are relatively stable and can be multiple. Any agent can be linked to several others, and at each turn may interact with one or many of their neighbors. The social networks can be static or evolve during a simulation, but the connections are typically more stable than the stochastic pairwise encounters. Social networks can model interesting social structures: see Chapter 3 for a longer discussion on the shape of real-life social networks and how these structures influence opinion dynamics.

Public diffusion corresponds to a social situation where agents can publicly broadcast their arguments to the community. This feature of communication is well suited to argumentative models, because pieces of information are explicitly represented as arguments. In contrast, models of opinion diffusion which simply assume a numerical opinion and no structure for the knowledge of agents rarely implement this mode of communication. Sometimes, all agents have access to the public arguments. In other cases, to guarantee that the mental argumentation frameworks of the agents remain distinct, there is a certain probability that an agent will perceive an argument. This mode of social interaction corresponds to situations where the agents have access to a way of sharing information openly: for instance, scientific communities (through article publication), decision committees or other forms of public deliberations.

In the following, I detail how each of the models implement these general principles in their social modeling.

[Gabbriellini and Torroni \(2014\)](#) implement an explicit social network. In this work, the population of agents is divided into highly connected clusters, with the optional presence of links between individuals of different clusters, the "bridges". Each agent selects a neighbor to interact with, with a higher probability of selecting

a member of their cave. This representation is based on Granovetter's theory of weak ties ([Granovetter, 1973](#)), which states that the stronger two people are linked, the stronger the overlap between their social networks, and thus most agents are part of strongly connected clusters. The weak ties are the "bridges" which connect clusters together. One of the goal of the article is to test whether the presence of these weak ties lowers polarization in the population.

The model of [Borg et al. \(2017a\)](#) uses a hybrid approach, with a social network and public diffusion of the arguments. First, the agents are placed within groups called collaborative networks, which may be homogeneous (composed of agents studying the same theory) or heterogeneous. Every five steps, agents share all of their information with other members of their collaborative network. In addition, the model implements a form of public diffusion: the agents share their information with any other agent with a certain probability. An interesting feature of this model is that the authors implement a form of bias in the way the agents share information: biased agents do not share the attacks that they have discovered towards their own theory. This mode of information sharing mimics the way research results are disseminated, both in close collaborative networks and publicly, with a certain amount of bias against negative results (which may be less likely to be published).

An underlying research question for [Taillandier et al. \(2021\)](#) is to investigate the effect of *homophily* on polarization. They adopt the same selection method as [Mäs and Flache \(2013\)](#) (see Chapter 3): partners for pairwise interactions are selected stochastically, with a higher probability of pairing two agents when their opinions are closer. [Ledoyen et al. \(2021\)](#) also use dyadic interactions ruled by homophily. They use the same process as [Deffuant et al. \(2000\)](#) to compute the similarity between two agent's opinions.

The interaction in [Dupuis de Tarlé et al. \(2022a\)](#) is public diffusion: all the agents interact together, and collectively participate in a public debate. They advance arguments which are added to a public argumentation framework. The agents learn the new arguments with a certain probability. The authors use this probability to model confirmation bias: the agents have a higher probability of learning arguments which favor their own opinion.

[Rago et al. \(2023\)](#) perform simulations with only two agents: a "human" and a "machine", in line with their applicative context, explainable artificial intelligence. These two agents always communicate with each other.

Finally, the model of [Butler et al. \(2019\)](#) implements two distinct social models: pairwise interactions and public diffusion. In pairwise interactions agents meet each other randomly. During the deliberative phases, a subset of agents participates in a collective debate, and advances arguments publicly.

4.3 Results

The results of these models shed light on the impact of argumentative discussion on the dynamics of agent's opinions. Many of these dynamics are more complex than those that emerge from non-argumentative opinion dynamic models.

[Gabbriellini and Torroni \(2014\)](#) show that some argumentation frameworks are "stronger" than others and as a result propagate more successfully in the population, in contrast to the classical take on opinions being of equal worth. As a result, they find that a small minority of agents can succeed in spreading their opinion to the majority.

The main result of [Dupuis de Tarlé et al. \(2022a\)](#) is that communication between agents does not always lead to a convergence of their opinions, unlike what is assumed in bounded-confidence models.

Similarly, [Butler et al. \(2019\)](#) find that deliberation causes more variance of opinion within the agents as well as a higher number of extremists. Their result coincide with some empirical findings that deliberation may polarize groups.

[Taillandier et al. \(2021\)](#) apply their model to a real-life case: using survey data, they study the diffusion of a favourable opinion of vegetarian diet in a population. Interestingly, they also report an increase in the number of agents with extreme opinions, and of the level of polarization. This suggests that argumentative discussion may be an important factor explaining the polarization of a population: this aspect is explored by some of the models described in Chapter 3.

[Ledoyen et al. \(2021\)](#) also apply their model to a real-life use case. They study the adoption of a new mechanism for irrigation in the South-West of France. They find that argumentation leads to a greater acceptance of the innovation and that introducing a new favorable argument for up to 50% of the population increases the number of agents which are favorable to the innovation, but that increasing this proportion has no noticeable effect.

Finally, [Borg et al. \(2017a\)](#) investigate a common question in ABMs of scientific communities: how different social networks impact the capacity of scientists to select the best theory. They find that an increased communication between the agents tends to benefit their epistemic success, which stands in opposition to the conclusions drawn from previous ABMs of scientific communities, such as Zollman networks (see Chapter 3).

4.4 Conclusion

As we have seen, there exists a growing literature of agent-based models which use abstract argumentation to represent the cognition of agents and the circulation of arguments within a group. Each of these models tackle a different research question, and adapts abstract argumentation to model specific social situation, whether it is collective decision making, scientific inquiry or innovation diffusion. This highlights the versatility of abstract argumentation, and the diversity of argumentation-related social contexts.

ArgABM (Borg et al., 2017a) is the only argumentative model which explicitly focuses on scientific debates. As such, it implement distinct mechanisms, such as a discovery relation between arguments. In the next section, I will present agent-based models of scientists which explore very similar research questions. One major difference is that my models do not use a pre-existing argumentative landscape: rather, in order to emulate the models of socio-technical controversies of O'Connor and Weatherall (2019), the agents generate new arguments at every turn. I built from the model of Dupuis de Tarlé et al. (2022a), introducing novel features such as bipolar graphs, a novel semantics, and additionnal mechanisms to represent a scientific debate (in Chapters 5 and 6). I also collaborated with the authors of Borg et al. (2017a) to propose a model of scientific debate which includes a form of cognitive bias known as myside bias (in Chapter 7). Thus, the models of the following section are most closely related to Dupuis de Tarlé et al. (2022a) for the structures and protocol, and Borg et al. (2017a) for the research questions.

Part II

Modelling Scientific Controversies with Abstract Argumentation

Introduction to Part II

Agent-based models allow us to simulate complex social interactions and test hypotheses in a controlled environment. They offer a promising approach to understand socio-technical controversies, such as those surrounding Covid vaccines, glyphosate, or global warming. A common and recommended method when developing complex ABMs with many different actors is to first build a simple model, validate it, and then gradually increase its complexity. For instance, researchers like Zollman and O'Connor, who work with Bayesian models, began by constructing and testing models focused on debates among scientists (Zollman, 2007; O'Connor and Weatherall, 2018). Later, they incorporated other types of actors to better understand and model the intricacies of socio-technical controversies (O'Connor and Weatherall, 2019). I followed a similar approach: my initial work involved building a robust model of debates within a scientific community.

While developing this model, building on my earlier work in Dupuis de Tarlé et al. (2022a), I realized that the gradual semantics I was using to assess the strength of arguments and the agents' opinions lacked certain desirable properties. By modifying this semantics, I created a new version with properties better suited for modeling agent opinions in an agent-based model. I published this novel semantics independently from the ABM, along with an analysis of its relevant properties, in Dupuis de Tarlé (2023). This work is detailed in Chapter 5.

Using this novel semantics, we created the model presented in Chapter 6. Inspired by the Bayesian models, it includes the generation of arguments by the agents, and is, to the best of my knowledge, the first to do so. After a first version of the model which was published in French in the proceedings of the 30èmes Journées Francophones sur les Systèmes Multi-Agents (JFSMA 2022) (Dupuis de Tarlé et al., 2022b), I improved the argument generation technique and created the model presented in Chapter 6, which was published as Dupuis de Tarlé et al. (2023) the 18th Social Simulation Conference and received the Best Student Paper Award. In this work, drawing from the existing literature in computational social epistemology, I investigated the role of **communication** on the epistemic success of the agents. Indeed, previous models of scientific communities such as Zollman (2007) and Hong and Page (2004) have shown that, surprisingly, communities where the agents are less connected and thus communicate less are better at selecting the best alternative. By testing the impact of communication on epistemic success, my work contributes to the discussion opened by these models. In Dupuis de Tarlé et al. (2022a), we had obtained the result that the exchange of arguments between agents could lead to their opinions becoming more distinct, so I knew that, unlike what is the case in Bayesian models, communication in argumentative models may not automatically result in the convergence of the opinion of agents. Thus, my question was larger: I investigated the relation between communication and opinion diversity and showed that argumentation leads to more complex dynamics than other opinion models.

Discussions with Matteo Michellini, a philosophy PhD student working on very similar research questions (and supervised by the authors of [Borg et al. \(2017a\)](#), the only previous argumentative ABM of a scientific community), led us to devise a model aimed at testing the theories of Mercier and Sperber about scientific reasoning. Mercier and Sperber are two cognitive scientists whose book *The Enigma of Reason* emphasizes the role of argumentation in human reasoning ([Mercier and Sperber, 2017](#)). They argue that all human reasoning is subject to **myside bias**—the tendency to produce arguments favoring one’s own point of view. This makes solitary reasoning prone to errors, as our minds generate more arguments that reinforce pre-existing beliefs. However, they propose that scientific reasoning avoids the negative effects of myside bias because scientific communities engage in collective discussions, allowing for the refutation of weak arguments. Additionally, the presence of ‘shared beliefs’ in scientific communities helps filter out weaker arguments and fosters consensus on what is considered valid. This hypothesis is well-suited for exploration through a formal argumentative model, as formalization forces precision in defining concepts and establishing clear boundaries between them. Moreover, the question of bias and shared beliefs was directly relevant to my investigations of socio-technical controversies. Understanding these dynamics provides deeper insights into how debates evolve, and how social structures shape the production, circulation, and eventual outcomes of arguments. In particular, the composition of groups at the outset and the presence of community-based filters proved to be highly influential in determining the trajectory of the controversy. This work was published, with Matteo and I as first authors, in the *Journal of Artificial Societies and Social Simulation* as [Dupuis de Tarlé et al. \(2024\)](#), and is presented in Chapter 7.

The two models of Chapter 6 and 7 were developed largely in parallel. This allowed me to compare their protocols as I constructed them. Thus, each model nourished the other whilst staying distinct, investigating separate research questions. A comparison of the two models and their results is presented in Chapter 8.

A Gradual Semantics to Model Opinion with Bipolar Argumentation Frameworks

This chapter introduces a novel opinion model and gradual semantics. It was published in the proceedings of the JIAF-JFPDA conference (Dupuis de Tarlé, 2023). This model and semantics are applied to the study of an epistemic community in the following Chapter 6.

5.1 Introduction

As stated in the previous chapter, the simplicity and flexibility of abstract argumentation makes it an ideal tool for the representation of knowledge, and the construction of multi-agent protocols. In particular, the inherent expressiveness of **gradual semantics** (see Chapter 2) makes them highly valuable for the development of multi-agent protocols, as they enable precise analysis of argumentative dynamics. For instance, in Dupuis de Tarlé et al. (2022a), we created a multi-agent model where the opinion of the agents is a number obtained through the application of a gradual semantic. This allows us to study the dynamics of the agents' opinions.

The first works on abstract argumentation consider only one relation between the arguments: the attack relation (Dung, 1995). Bipolar graphs are an extension of classical abstract argumentation graphs which consider an additional relation, that of *support* (Cayrol and Lagasquie-Schiex, 2005). These graphs are more expressive and have been validated by empirical experiments as more representative of the way in which humans actually reason (Polberg and Hunter, 2018). This work proposes to extend the analysis of Dupuis de Tarlé et al. (2022a) to the case of bipolar argumentation graphs.

Generalizing from the model of Dupuis de Tarlé et al. (2022a), we build an opinion framework which enables the representation of agents' knowledge in the form of a bipolar graph, where their opinion is obtained by applying a gradual semantic. As developed in Chapter 4, Rago et al. (2023) propose a similar general opinion framework and applies it to the problem of AI explainability: a machine and a human agent use argumentative exchange to resolve conflicts and generate explanations. Their work, which uses bipolar argumentation graphs and gradual semantics, is quite close to ours, but the application to explainable AI is distinct from that of opinion dynamics, and they do not feature any discussion about the properties of gradual

semantics. Their work and my own were developed independently and at the same time, which indicates the relevance of extending the protocol of Dupuis de Tarlé et al. (2022a) with bipolar graphs. The studies of properties of gradual semantics sometimes lack discussion about their relevance to specific use contexts (Amgoud et al., 2017; Potyka, 2018). We propose to bridge the gap between theory and application by defining a general opinion framework, which leads us to identify some desirable properties of such gradual semantics: in particular, the properties of open-mindedness and duality identified by Potyka (Potyka, 2018, 2019). We propose a new semantics which satisfies these principles. The novel contributions are the opinion model and the gradual semantics, while the properties have already been proposed in the literature.

The first section of this chapter introduces a general framework that effectively represents agents' opinions using argumentation graphs. This framework serves as a foundation for the presentation of the various desirable properties of semantics which have been proposed in the literature in the subsequent second section. To address these properties, the third section introduces a novel gradual semantics that ensures, among others, the presence of open-mindedness and duality. Moreover, we delve into the implications of applying this semantics to our opinion framework within this section.

5.2 Opinion Model

In this section, we place ourselves within the context of an argumentative discussion between agents, and we define a way to characterize the opinion of the agents through a bipolar graph, their **opinion graph**. We do not specify a multi-agent protocol governing what actions are performed each turn, so we use the terminology "framework" or "structure". Our goal is to create a flexible model that can serve as a basis for the creation of various multi-agent protocols, whose specific characteristics would make it possible to study various phenomena.

The opinion graph can be interpreted either as the agent's knowledge base, or the arguments that she takes into account in her evaluation of a debate. Here, we favor the first interpretation, and say that an agent *knows* an argument when it belongs to her opinion graph.

Following the methodology of Bonzon and Maudet (2012) and Dupuis de Tarlé et al. (2022a), all of the opinion graphs contain an argument with a special status, the *issue*, which constitutes the focus of the debate. These graphs are **issue-oriented**, which means that all the arguments of a graph belong to a path of supports and attacks directed towards the issue. We interpret the acceptability score of the issue given by the semantics and the agents' graphs as the agent's **opinion**. Thus, the opinion of the agents is a real number, which belongs to the image set of the gradual semantics that we use (in most case, this interval is $[0, 1]$). For instance, an agent

could have the opinion graph represented in Figure 2.5, with argument a as the issue, and this agent's opinion would be 0.53.

We represent the opinion of the agent as a degree of acceptability about a single issue. This is warranted by the context of argumentative discussions, as we make the hypothesis that the discussion is focused on determining the acceptability of a single question, for instance in a debate about the proposition to make all agriculture organic. However, this framework could be extended to include multidimensional opinions about several issues.

Many seminal opinion dynamics models represent the opinion of agents as a real number in the interval $[0, 1]$: such is the case of the *bounded-confidence* models (Hegselmann and Krause, 2002; Deffuant et al., 2000). These models make the assumption that the opinion of agents can be represented as a real number for the sake of simplicity, citing the example of "an expert who has to assess a certain magnitude" (Hegselmann and Krause, 2002). In our case, as we are studying argumentative discussions, one natural interpretation of the opinion is a **degree of belief** of the agent in the acceptability of the issue.

We can now formally define our framework, starting with opinion graphs.

Definition 9 (Issue-Oriented Bipolar Graphs). *Let $B = \langle A, R, S, W \rangle$ a weighted bipolar graph. B is **issue-oriented** if there exists $i \in A$ such that for all $a \in A$, there is a path from a to i .*

Suppose we have a semantics for weighted bipolar graphs $\sigma : G \times A \rightarrow D \in \mathbb{R}$. We can then define the opinion of agents as the evaluation of the acceptability of the issue by the gradual semantics σ applied to their opinion graph.

Definition 10 (Opinion of an agent). *Let k be an agent equipped with an issue-oriented bipolar graph of issue i , $B_k = \langle A, R, S, W \rangle \in G$ and $\sigma : G \times A \rightarrow D \in \mathbb{R}$ a modular gradual semantics well defined on B_k . We define the **opinion** of agent k as $O_k = \sigma(B_k, i)$.*

The main contribution of our framework is that the agents' opinions are derived from graphs forming their knowledge base, using gradual semantics which can express certain ideals of rationality. With an accurate choice of image set D for the semantics as $[0, 1]$, the results of multi-agent protocols built using our framework could be directly compared with that of the bounded-confidence type models.

We can see that the semantics used plays a major role in the evaluation of the agents' opinions. The dynamics of the agent's opinions will be highly influenced by the properties of the semantics. In the next section, we delve into desirable properties for this context.

5.3 Desirable Properties

5.3.1 First principles

[Amgoud and Ben-Naim \(2018\)](#) carry out an extensive study of gradual semantics for weighted bipolar graphs. The authors identify twelve desirable properties that can be satisfied by such semantics. Table 5.1 offers an intuitive explanation of each of these principles. We refer the reader to the original article for a complete formalization.

The authors compare existing semantics for weighted bipolar graphs based on these principles. They propose the *Euler-based semantic* (EBS) and show that it is the only semantics that satisfies their twelve principles.

Let us define formally the Euler-based semantics introduced by [Amgoud and Ben-Naim \(2018\)](#). This semantics is defined exclusively in the case of acyclic bipolar graphs and is based on a quantity, the energy, which aggregates the scores of the supports and direct attackers of an argument.

Definition 11 (Energy of an argument). *Let $B = \langle A, R, S, W \rangle \in G$ a weighted bipolar graph. For an argument $a \in A$, $\text{Supp}(B, a)$ and $\text{Att}(B, a)$ are respectively the set of arguments supporting and attacking a in B . The energy E is defined as the function $E : G \times A \rightarrow \mathbb{R}$ such that for all $a \in A$:*

$$E(B, a) = \sum_{x \in \text{Supp}(B, a)} \sigma(x) - \sum_{x \in \text{Att}(B, a)} \sigma(x)$$

Definition 12 (Euler-Based Semantic (EBS)). *Let $B = \langle A, R, S, W \rangle \in G$ an acyclic weighted bipolar graph. $\text{Ebs}(B)$ is the score function $\sigma : G \times A \rightarrow [0, 1]$ recursively defined as : For all $a \in A$ of weight w_a*

$$\sigma(B, a) = 1 - \frac{1 - w_a^2}{1 + w_a e^{E(B, a)}} \quad (5.1)$$

As noted in [Potyka \(2019\)](#), the acceptability score of an argument evaluated by EBS cannot be less than the square of the weight of the argument. Figure 5.1 represents the variation of the score of an argument according to EBS against its energy $E(B, a)$, in the case $w_{base} = 0.5$. We see that this score varies between 0.25 and 1.

Table 5.1: Principles defined by Amgoud and Ben-Naim (2018) which can be satisfied by a gradual semantics for weighted bipolar graphs.

Property	Intuition
Anonymity	The score of an argument is independant from its identity.
Bivariate Independance	The score of an argument is independant from every argument that is not linked to it.
Bivariate Directionality	Only the relations directed towards the argument influence its score, and not relations directed away from it.
Bivariate Equivalence	The score of an argument only depends on its weight and on the score of its direct attackers and supporters.
Stability	If an argument is neither attacked nor supported, its score must be equal to its weight.
Neutrality	Attackers and supporters of score equal to zero have no effect on their targets.
Bivariate Monotony	If an argument a is as much or less attacked than an argument b , and as much or less supported than b , then the score of a must be at least as great as that of b .
Bivariate Reinforcement	An argument's score increases if the quality of its attackers is reduced and the quality of its supports is increased.
Resilience	If an argument's weight is positive, its score cannot be reduced to zero by attacks. If the weight is lower than 1, it cannot reach 1 with supports.
Strict Franklin	Attacks are as important as supports.
Weakening / Strengthening	If attacks are greater than supports, the score of the argument is lower than its weight, and conversely.

5.3.2 EBS and the opinion model

If we use the EBS semantics in the context of the model for representing agents' knowledge and opinions presented in Section 5.2, we need to define a **weight** w_i for the issue. The behavior of EBS means that the opinion of the agents would all be comprised in the interval $[w_i^2, 1]$. Among other things, this limits the possibilities of comparison with *bounded confidence* type models. While this may be a reasonable behavior if one interprets the weight of the issue as its intrinsic strength, it would be less justifiable in a protocol *without weights*.

Indeed, most of the gradual semantics for bipolar graphs proposed in the literature are defined for weighted graphs. The underlying logic is that of "Whoever can

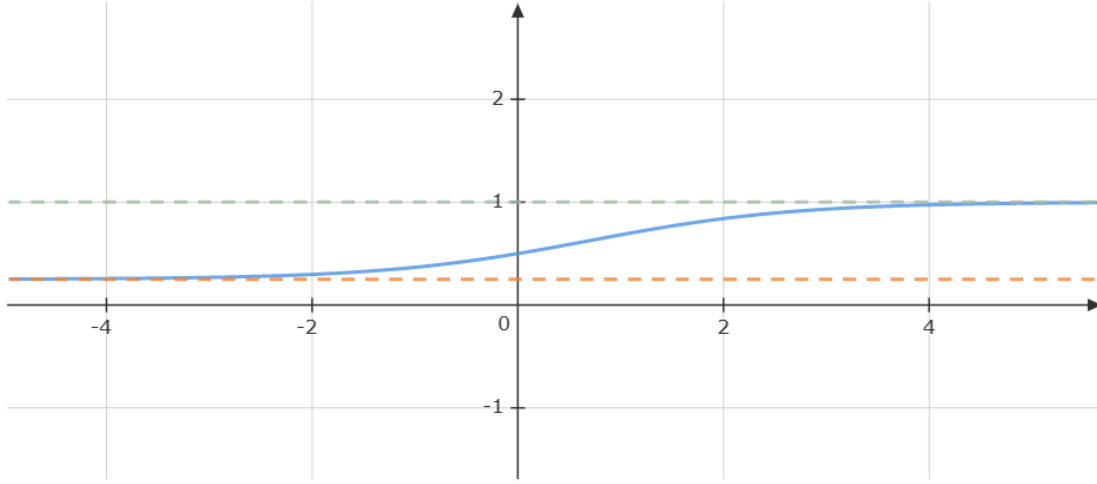


Figure 5.1: Variation of the score of an argument according to EBS (y-axis) against its energy $E(B, a)$ (x-axis), in the case $w_{base} = 0.5$. We see that the value of the score is between 1 (grey line) and $w_{base}^2 = 0.25$ (orange line).

do the most can do the least", i.e. a semantics capable of accommodating an additional level of complexity can *a fortiori* deal with simpler cases, here non-weighted graphs. It would suffice to choose the weight well to obtain a semantics that behaves correctly. This is a method successfully applied for attack (and support) gradual semantics: thus, the h-categorizer (Besnard and Hunter, 2001) semantics can be adapted to non-weighted attack graphs and retains desirable properties (Potyka, 2019). This is not the case with EBS, as was already noticed in Cocarascu et al. (2019).

If one wishes to apply EBS to non-weighted graphs, it is necessary to choose a weight w_{base} for all the arguments, which will correspond to their evaluation when they are neither attacked nor supported (a natural choice for such a weight would be 0.5 for example). However, because the acceptability score of an argument evaluated by EBS cannot be less than the square of the weight of the argument, this semantics adapted to a non-weighted graph would no longer affect a score between 0 and 1 but in the interval $[w_{base}^2, 1]$. Figure 5.1 illustrates this situation in the case of $w_{base} = 0.5$. This would limit the opinions of the agents to the interval $[w_{base}^2, 1]$, without any clear justification ¹.

Another consequence of this behavior is that, because the range of value $[w_i, 1]$ is greater than $[w_i^2, w_i]$, the impacts of supports on the opinion of the agents are proportionally greater than those of attacks, regardless of the weight of the argument.

These two behaviors are related to the properties of open-mindedness and duality.

¹Note that by choosing a weight of 0, we would ensure an opinion interval of $[0, 1]$ but that in this case, all the scores of the arguments would be equal to 0.

5.3.3 Open Mindedness and Duality

Potyka's works (Potyka, 2019, 2018) define two other properties of gradual semantics for bipolar graphs: open-mindedness and duality.

Intuitively, open mindedness corresponds to the fact that the score of an argument can vary freely between the limits of interval D , whatever its weight: given enough attacks and supports, its score may attain any number between 0 or 1.

Definition 13 (Open Mindedness). *Let $\sigma : G \times A \rightarrow D$ a semantics for weighted bipolar graphs on an interval D . The semantics is open-minded for all graph $B = \langle A, R, S, W \rangle$ if for all argument $a \in A$ and all $\epsilon > 0$, the following condition is satisfied : there exists a number $N \in \mathbb{N}$ such that if we add N new arguments whose base score is maximal : $A_N = a_1, \dots, a_N$, $A \cap A_N = \emptyset$, then :*

1. *For graph $B_{att} = \langle A \cup A_N, R \cup \{(a_i, a) | 1 \leq i \leq N\}, S, W' \rangle$, we obtain $|\sigma(B_{att}, a) - \min(D)| < \epsilon$*
2. *For graph $B_{supp} = \langle A \cup A_N, S \cup \{(a_i, a) | 1 \leq i \leq N\}, S, W' \rangle$, we obtain $|\sigma(B_{supp}, a) - \max(D)| < \epsilon$*

where $W'(b) = W(b)$ for all $b \in A$ and $W'(a_i) = \max(D)$ for $i \in \llbracket 1, N \rrbracket$.

Potyka (2018) also defines a property which illustrates the intuitive notion of symmetry between the actions of attacks and supports: duality. To illustrate, let us take the example of EBS: the asymmetric nature of this semantics causes an imbalance between the action of attacks and supports. By "symmetrical action", we mean that if the attackers influence the score such that they modify the weight of the arguments by removing a certain quantity, turning them into supporters should have the inverse effect, that is to increase the score by the same quantity. One would expect symmetrical actions of attacks and supports when the weight is 0.5. To illustrate this, let us consider argument l of weight $w_l = 0.5$, attacked by argument m . If the score of l according to some semantics is $0.5 - 0.1 = 0.4$, then a desirable property would be that, if we were to turn the relation from m to l into a support, the score of l would be $0.5 + 0.1 = 0.6$. On the other hand, when the weight is greater or less than 0.5, this symmetry requirement is more restrictive because the weight of the argument is now closer to one of the limits of the interval $[0, 1]$. For instance, the weight of l is $w_l = 0.95$, then any action of a group of supports cannot exceed 0.05, as the score cannot be greater than 1. Therefore, the symmetry requirement would enforce that no (set of) attackers may decrease the score below 0.9. As such, symmetry may be less desirable as it would not be compatible with open-mindedness.

To avoid this problem, (Potyka, 2018) generalizes the symmetry intuition to all weights in the following way: suppose that the initial weights of two arguments x

and y are shifted relative to 0.5 by ϵ in different directions, and that the attackers of x have the same strength as the supports of y and vice versa. Then if the application of a dual semantics to x transforms its weight into a score shifted by δ , the score of y should be shifted by $-\delta$ with respect to its weight. We define this property, duality, in the context of modular gradual semantics.

Definition 14 (Duality). *Let $\sigma : G \times A \rightarrow D$ be a modular semantics for weighted bipolar graphs, with α its aggregation function and ι its influence function. The semantics σ satisfies **duality** for all $B = \langle A, R, S, W \rangle$ if and only if it satisfies the following property :*

Let $a, b \in A$ such that $w_a = 0.5 + \epsilon, w_b = 0.5 - \epsilon$ for an $\epsilon \in [0, 0.5[$, and the supporters and attackers of a and b are such that :

$$\begin{aligned}\alpha(B \setminus \text{Att}(B, a), a) &= -\alpha(B \setminus \text{Supp}(B, b), b) \\ \alpha(B \setminus \text{Att}(B, b), b) &= -\alpha(B \setminus \text{Supp}(B, a), a)\end{aligned}$$

then $\sigma(B, a) - w_a = w_b - \sigma(B, b)$.

5.3.3.1 Influence of open mindedness and duality in different contexts

Concerning open-mindedness: apart from not being very elegant, real problems can emerge from a situation where the entirety of the opinion space is not accessible to agents. For instance, in Chapter 6 we create an argumentative model studying epistemic communities where, like in Hegselmann's work (Hegselmann et al., 2006), the success of the agents is measured by a distance between their opinion and a truth value. In this case, it is very important that the opinion of the agents can approach any value of the interval $[0, 1]$. Similar problems would arise if one wanted to study extremism, another phenomenon investigated in bounded-confidence type models (Deffuant et al., 2002).

Duality, on the other hand, imposes a form of symmetry between the actions of attacks and supports. If it is not satisfied, the dynamics may differ from one side of the opinion space to another, which could be problematic for certain protocols. For instance, Zhong and Negre (2022) create an argumentative recommendation system. The program recommends an item to a user by aggregating pro and con arguments. This work could be extended by using a gradual semantics; and in this case, a symmetry between the influence of the supports and the attacks may be desirable. Otherwise, the system would be biased to assign more importance to certain types of arguments, which may not be justifiable. Rago et al. (2023) present an application which can also be modelled within our framework. The authors use an opinion model with gradual semantics to model explainability: a machine and a human resolve a conflict by exchanging arguments. In this context, it may be interesting to allow the opinion of the human or machine agent to explore the full

interval $[0, 1]$. Similarly, there would be no reason to treat supports as having more influence than attacks, or conversely.

This is not to say that duality or open mindedness are always preferable, in all contexts. Rather, we want to shed light on the fact that properties such as these should be considered and chosen within an application context, rather than imposed because of the lack of an alternative.

EBS satisfies neither open-mindedness, nor duality. In the following section, we propose a novel semantics which satisfies both of these properties.

5.4 Novel Semantics

We define a modular gradual semantics by combining the *energy* aggregation function with a modified logistic influence function. Like EBS, our semantics is defined exclusively for acyclic graphs.

Definition 15 (*Logistic Sum Semantics (LSS)*). *Let $B = \langle A, R, S, W \rangle$ be an acyclic bipolar graph. LSS is the score function $\sigma : G \times A \rightarrow [0, 1]$ recursively defined by : For all $a \in A$ of weight w_a , if $w_a = 0$, $\sigma(B, a) = 0$, if $w_a = 1$, $\sigma(B, a) = 1$ and if $w_a \in]0, 1[$:*

$$\sigma(B, a) = 1 - \frac{1}{1 + e^{E(B, a) + c(w_a)}}, \quad (5.2)$$

with $c(w_a) = \ln(\frac{1}{1-w_a} - 1)$.

The behavior of the LSS semantics in the boundary cases of $w_a \in \{0, 1\}$, is the same as EBS: in these cases, adding attacks or supports does not modify the score of the argument. These cases were included to allow certain manipulations, such as adding attacks and supports of score 1 or 0.

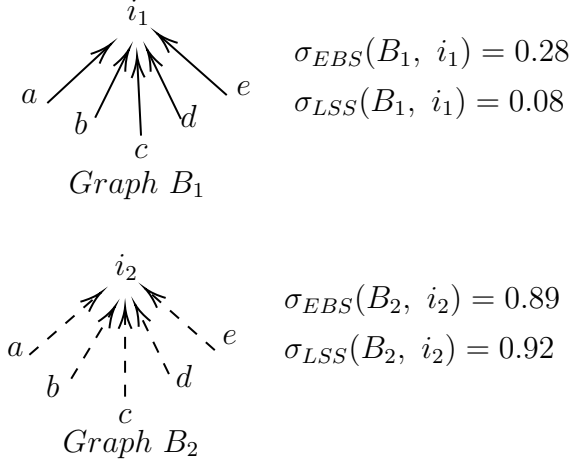
Property 1. *The LSS semantics satisfies the twelve principles defined by (Amgoud and Ben-Naim, 2018) (see Table 5.1).*

Property 2. *The LSS semantics satisfies open mindedness and duality.*

Technically, it is the restriction of the LSS semantics on arguments of weight in $]0, 1[$ which satisfies these desirable properties. Indeed, since the score of arguments in the boundary cases, with a weight of 0 or 1, cannot be modified by their attackers and supporters, open mindedness, duality, or some of the properties described in Table 5.1 like balance and monotonicity cannot be satisfied. However, as noted by

(Potyka, 2019), EBS does not verify balance and monotonicity in the the boundary cases either. The proof of Property 5.4 is discussed in more details in the Appendix.

The following example compares the behavior of LSS and EBS on two simple argumentation graphs, and illustrates open-mindedness and duality.



Consider the above graphs B_1 and B_2 , where we use full arrows to indicate attack relations and dashed arrows to indicate supports. The issue of B_1 is attacked by 5 arguments, and the issue of B_2 is supported by 5 arguments. We fix all of the weights of the arguments at 0.5, thus the aggregation of the attackers of i_1 is equal to the inverse of the aggregation of the supporters of i_2 (using the energy function).

The example illustrates the problem mentioned above, which is that EBS is limited to the interval $[0.25, 1]$. Indeed, the value of i_1 is 0.28 according to EBS, while LSS is able to assign a lower value of 0.08. We can also note that if we were to add attacks to i_1 , EBS would not show much modification because the score of i_1 is already close to the limit of 0.25, while LSS would be more expressive, but on the other hand their evaluations of i_2 are much more similar.

In this context, duality is satisfied if the sum of the evaluation of i_1 and i_2 is equal to 1, and we can see that it is satisfied by LSS and not by EBS.

Another illustration of these two properties in the case of LSS can be found on Figure 5.2, where we represent the score of an argument against its energy for three different weights. We see that in all three cases, open mindedness is satisfied as the score covers the whole interval. We can see that the purple and the pink curve, which corresponds to the acceptability score of an argument with a weight of 0.8 and 0.2 respectively, exhibit a central symmetry around the point $(0, 0.5)$. This symmetry corresponds to duality: if we take two arguments a and b with $w_a = 0.8$ and $w_b = 0.2$, and if their energies are opposite from each other, then the sum of their acceptability scores will be equal to the sum of their weights, which is 1.

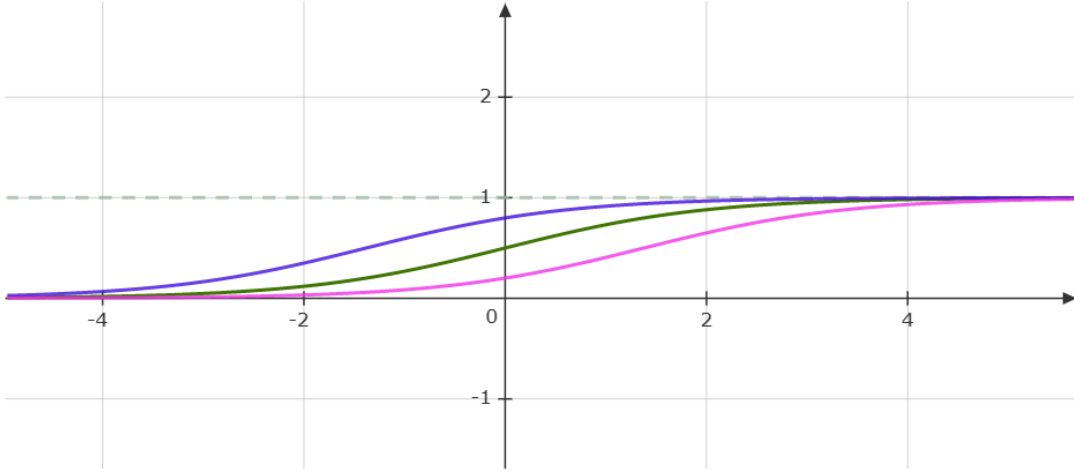


Figure 5.2: Variation of the score of an argument according to LSS (y-axis) against its energy $E(B, a)$ (x-axis), for three weights $w_a = 0.2, 0.5$ and 0.8 respectively in pink, green and blue. We see that the value of the score is between 1 (grey line) and 0 and that when $E(B, a) = 0$, the score is equal to the weight.

5.4.1 Convergence of Opinions

Let us now place ourselves in the general framework defined in Section 5.2: consider agents equipped with an opinion graph and let us use the LSS semantic. Suppose that two agents communicate by exchange of arguments, what can be said about their respective opinions?

This question is important because it will allow us to compare directly any protocol built with our framework to the bounded confidence type models, where any communication between agents automatically results in a convergence of their opinions.

Dupuis de Tarlé et al. (2022a) show that in the case of attack graphs, with the h-categorizer semantic, the communication between agents does not automatically result in a convergence of their opinions. However, their simulations show an empiric convergence when many interactions take place.

In order to study this problem, we need to formally define what we mean by communication through exchange of arguments. For this, we make a number of simplification assumptions.

- The opinion graphs of the agents are acyclic.
- When agents are aware of the same arguments, they are also aware of the same attack and support relations between them.
- Agents all agree on the weights of their shared arguments.

The initial assumption enables the use of our LSS semantic. The other two assumptions, which are aligned with, for instance, the protocol defined in [Dupuis de Tarlé et al. \(2022a\)](#), are rather restrictive and allow us to define communication as a strict exchange of arguments without requiring a merging mechanism for attacks, supports, and weights. It is worth noting that these constraints, within which a wide variety of protocols can still be constructed, could be relaxed within our framework given that we ensure the existence of a compatible semantics and establish a process for merging argumentation graphs.

In accordance with the idea that opinion graphs are agents' knowledge bases, we say that an agent *learns* an argument when she adds it to her opinion graph.

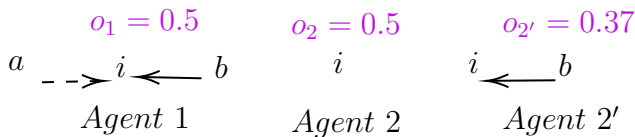
Definition 16 (Learning an argument). *Let an agent k equipped with opinion graph $B_k = \langle A, R, S, W \rangle$, and (a, R_a, S_a, w_a) a tuple composed of an argument a , relations R_a and S_a such that $R_a = \{(a, x) | x \in A_0 \subset A\} \cup \{(x, a) | x \in A_1 \subset A\}$ and $S_a = \{(a, x) | x \in A_2 \subset A\} \cup \{(x, a) | x \in A_3 \subset A\}$, and a weight $w_a \in [0, 1]$. Agent k **learns** argument a by transforming her opinion graph to $B'_k = \langle A \cup \{a\}, R \cup R_a, S \cup S_a, W' \rangle$ with $\forall x \in A, W' : x \rightarrow W(x)$ and $W'(a) = w_a$.*

We suppose that the attack and support relations R_a and S_a that link argument a to the arguments of the opinion graph of agent k are known. Depending on the specifics of the protocol, they could be obtained from another agent's opinion graph, or generated dynamically. Thus, we can define communication between agents as the learning of arguments from other's opinion graphs.

Under these constraints, we show the following property.

Property 3. *When using the LSS semantic, the opinion of two agents does not necessarily converge when they exchange arguments from their opinion graphs.*

It is easy to generate an example that illustrates (and proves) Property 3. Consider two agents 1 and 2 whose opinion graphs are shown below.



When applying the LSS semantics on their initial opinion graphs, their opinions are the same: $o_1 = o_2 = 0.5$. Consider what happens if Agent 2 adds one of the arguments of Agent 1 to her opinion graph: this is the situation denoted Agent 2'

above. Her opinion $o_{2'} = 0.37$ is now further from that of Agent 1, even though their opinion graphs are now more similar².

Property 3 is also satisfied when using the EBS semantic. This result supports the claim that gradual semantics exhibit non-trivial properties that justify the interest of their study in the context of multi-agent models.

5.5 Conclusion

We proposed a general structure that represents the knowledge and opinions of agents using **weighted bipolar argumentation graphs** and **gradual semantics**. This approach allows for the nuanced modeling of the dynamics of opinions in agent-based models. We also discussed various desirable properties for such semantics, particularly focusing on the principles of **open-mindedness** and **duality**, and proposed a new gradual semantics that satisfies these principles.

In the future, we plan to extend this work in several directions. First, we aim to verify whether our proposed gradual semantics, LSS, satisfies additional properties such as those described in Baroni et al. (2018) and Rago et al. (2018). We also plan to extend the comparison of our semantics beyond EBS to include other gradual semantics for bipolar graphs, such as QuAD and Df-QuAD Baroni et al. (2015); Rago et al. (2016), and the quadratic energy semantics proposed by Potyka (2018). This will help to establish a broader understanding of the performance and applicability of our proposed semantics in different contexts.

Through our model, we identified that the use of our gradual semantics gives rise to non-trivial opinion dynamics, in particular that agents' opinions do not necessarily converge when they communicate. Given similar results obtained by Dupuis de Tarlé et al. (2022a), we believe this behavior is not unique to our semantics. Therefore, a natural extension of this work would be to characterize a minimal set of properties that a gradual semantics must satisfy to guarantee this type of behavior.

We also plan to conduct a comparison study using an agent-based protocol to assess the impact of various gradual semantics on the dynamics of agents' opinions. By analyzing how different semantics influence the process of opinion formation and change, we hope to better understand which semantics may be most appropriate for different applications, such as online debate platforms or agent-based simulations of socio-technical controversies.

This future work will deepen our understanding of how to model complex debates and how to study opinion dynamics in environments characterized by multiple

²The notion of similarity between graphs is not developed here, but consider for example a graph edit distance, which defines distance between two graphs as the number of editing operations needed to transform one graph into the other (Sanfeliu and Fu, 1983).

interacting agents.

CHAPTER 6

Exploring Opinion Diversity and Epistemic Success in an Argumentative Model

This chapter was published in the proceedings of the 18th Social Simulation Conference, where it was awarded the Best Student Paper Award ([Dupuis de Tarlé et al., 2023](#)).

6.1 Introduction

Social epistemology is a branch of philosophy which explores the impact of social life and organisation on the pursuit of truth ([Goldman, 1999](#)). Agent-based simulation is a compelling tool for examining this relationship. Philosophers have recently used argumentation to understand the group dynamics of truth-seeking agents ([Mercier and Heintz, 2014](#)). By modelling argumentative processes, we aim to analyse and understand the effect of some features of group discussion on the epistemic outcome.

We present a model that depicts an **epistemic community**, a group of agents seeking to uncover the truth regarding a specific issue (scientific communities are typical examples of epistemic communities). The agents engage in a dynamic argumentative discussion to discover this truth. It is assumed that experience grants them access to certain facets of reality and that the discussion entails a repeated exchange of conflicting and supportive arguments. The assumption that reasoning is done through the exchange of arguments is in line with the argumentative theory of reasoning defended by [Mercier and Sperber \(2011\)](#). The arguments, organized in abstract argumentation frameworks and equipped with a novel gradual semantic, constitute the information base of the agents and can be shared between them in a mechanism of social influence.

Our model shows many similarities with opinion diffusion models: like in the seminal bounded confidence-type models presented in Chapter 3, agents' opinions are numbers between 0 and 1, and agents influence each other through pairwise interaction. One major novelty of our model is that such behavior is modelled via the explicit representation of information as an organized set of objects, the arguments, which can be generated and transmitted between agents.

Our model merges distinct areas of multi-agent simulation: computational social epistemology and argumentative agent-based models.

As is developed in Chapter 3, the foundations of social epistemology, which studies the knowledge and beliefs of groups of individuals, were introduced by Goldman (1999). Subsequently, many works have used multi-agent systems to study scientific communities as examples of epistemic communities. Zollman (2007) and O'Connor and Weatherall (2018) use models of Bayesian agent networks that interact with an environment and exchange the results of their experiments.

With the development of abstract argumentation as an effective way to represent debates as graphs, many agent-based models have been created to simulate argumentative discussions (see Chapter 4). Our work uses the general opinion framework presented in Chapter 5. We use the novel gradual semantics LSS to model the opinion of agents. We introduce a protocol, using the notions of graph generation and interaction between agents and the environment. Dupuis de Tarlé et al. (2022a) show that when agents learn each other's arguments, their opinions converge. Our model challenges these results, as we show that exchanging arguments can also lead to a greater diversity of opinions.

In our model, agents discuss the validity of a central question, the *issue*, and their opinions vary between 0 and 1. A major assumption of our model is that we assume that the *issue* and all subsequently generated arguments are characterized by a *truth value* in the interval $[0, 1]$. This way of modelling the truth has precedents: (Hegselmann et al., 2006) uses a similar truth value to extend the bounded confidence model to a community of truth-seekers. This is reminiscent of fuzzy logic, which uses continuous truth values interpreted as partial truths.

Our modelling approach is outlined in three sections. The first provides a brief reminder of the formal tools utilized: abstract argumentation and the gradual semantic we propose. The second section outlines our model and protocol. The third section delves into two crucial aspects of our model: the evaluation of epistemic success (i.e. the ability to approach the truth), and examination of opinion convergence. We show that the relation between the number of connections between agents and the diversity of their opinions is not monotonous, and, departing from the seminal results of Bayesian models, that opinion diversity is harmful for their epistemic success.

6.2 Abstract Argumentation

6.2.1 Bipolar argumentation frameworks

A bipolar abstract argumentation framework (BipoAF), formally $\mathbf{B} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$, consists of a finite set of arguments $\mathcal{A}$, equipped with two binary relations on this set: the attack relation $\mathcal{R}$ and the support relation $\mathcal{S}$ (Cayrol and Lagasque-Schiex, 2005). BipoAFs can be represented by graphs whose nodes are the arguments and

edges represent the relation between them. In the rest of this chapter, we will use the terms BipoAF and bipolar argumentation graphs independently, and often refer simply to “argumentation graphs”. Figure 6.1 presents a small debate made up of five arguments, and their representation as an argumentation graph with the relations that binds them: we used full lines to represent an attack, and dashed lines to represent support.

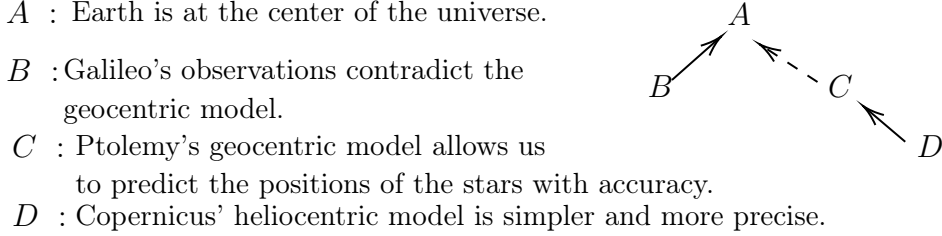


Figure 6.1: Example of a debate and its representation as a bipolar abstract argumentation graph.

6.2.2 Logistic sum semantic for bipolar graphs

In Chapter 5 we presented a novel gradual semantic for bipolar argumentation graphs, the **Logistic Sum** gradual semantic (LSS). This function, which we justify as being well adapted to represent the opinion of agents, verifies a number of desirable properties and assigns to every argument a **score** between 0 and 1. This semantic is defined for *weighted* bipolar argumentation graphs, where arguments are equipped with *weights*; numerical values which represent an intrinsic quality of the argument. We use the version of this semantics which is adapted for non-weighted graphs, by assuming a weight of 0.5 for all of our arguments.

Definition 17 (Non-weighted Logistic Sum Semantic). *Let $\mathbf{B} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be an acyclic bipolar graph and $a \in \mathcal{A}$ one of its arguments. LSS is the score function $\mathfrak{J}(a)$ recursively defined as :*

$$\sigma(B, a) = 1 - \frac{1}{1 + e^{E(B, a)}}, \quad E(B, a) = \sum_{x \in \text{Supp}(B, a)} \sigma(B, x) - \sum_{x \in \text{Att}(B, a)} \sigma(B, x) \quad (6.1)$$

with $\text{Supp}(a)$ and $\text{Att}(a)$ the sets of arguments which are respectively supporting and attacking a .

We see that arguments which are neither attacked nor supported, or whose attacks are compensated by supports, have a score of 0.5. The more attacked an argument is, the closer its score is to 0, and conversely, arguments that are well supported have a score that tends towards 1. This semantic thus allows us to derive, for each agent, an opinion about an issue which is a number between 0 and 1 based on what they know of the attack and support towards this issue. The following section details this mechanism.

6.3 The Model

6.3.1 Setting

In our model, agents generate arguments which form bipolar graphs. These BipAF, like those introduced in Chapter 5, are *issue*-oriented argumentation graphs. A special argument, the *issue*, is the root of the graph, and all other arguments are part of a path to this *issue*. Intuitively, the issue represents the central proposition of the debate, and all agents' moves are aimed at (indirectly) attacking or supporting that issue.

Each agent $k \in N$ is equipped with her own argumentation graph $\mathcal{O}_k = \langle \mathcal{A}_k, \mathcal{R}_k, \mathcal{S}_k \rangle$, which will contain the argument that she generates and those that she receives from her neighbours. This graph is called the **opinion graph** of the agent. The agents apply the non-weighted Euler based semantic to their opinion graph and derive a *score* $\mathfrak{Z}_k(a)$ for each argument $a \in \mathcal{A}_k$. The score of the issue is the **opinion** of the agent, o_k .

Each argument a is also equipped with an intrinsic **truth value** T_a . The agents cannot access these truth values. The truth value of the issue, T_i , is of particular interest to our model and is sometimes referred to as the truth value of the debate. We measure the epistemic success of the agents by how distant their opinion (assessment of the issue) is from T_i .

6.3.1.1 Network communication

Agents are placed in a **social network** $\mathcal{N}$, and share all the arguments that they produce with their direct neighbours. We use Erdos-Renyi non-directed random graphs (Erdős et al., 1960) to model this network: a network $\mathcal{N}$ is characterized by the number of agents N and p_{ER} , the probability that two agents are linked (see Chapter 3).

6.3.2 Protocol

At the beginning, all agents' opinion graphs are composed only of the issue. Therefore, the initial opinion of all agents is the same: $\forall k \in N, o_k = 0.5$.

Each step t of our protocol is composed of two main stages.

Stage 1 : Argument Generation

- Every agent k chooses randomly an argument a from her opinion graph.
- The agent investigates the chosen argument and, according to the process detailed below, generates a new argument¹ a' and either an attack or a support relation from a' towards a .

Stage 2 : Argument Sharing

- Every agent adds to their own opinion graph the argument and relation they generated.
- Every agent adds to their own opinion graph the argument and relations generated by their neighbours in $\mathcal{N}^2$.

This two-stage protocol allows us to simulate the simultaneous activation of the agents. Note that the opinion graph (and thus the opinion) of the agents changes due to their own generation of argument as well as that of their neighbors. This means that even in situations where agents do not communicate with each other, their opinions evolve.

6.3.2.1 Argument generation

The generation of attacks and supports is inspired by the one-armed bandit mechanisms used in Bayesian networks of agents (Zollman, 2007; Rosenstock et al., 2017).

In our argument generation method, agents sample a result from a distribution centered on the investigated argument's truth value. A natural interpretation is that the sampled result obtained corresponds to an "experiment" with the world. The result of this experiment is compared to the agent's previous perception of the investigated argument. If it is higher, the result has exceeded the agent's expectation, and the agent will produce a support for the investigated argument. Conversely, if it is lower, the agent will produce an attack of the investigated argument. The agents only have direct access to their sampled results and their own opinion.

Formally, let us consider agent k , whose opinion graph is $\mathcal{O}_k = \langle \mathcal{A}_k, \mathcal{R}_k, \mathcal{S}_k \rangle$. Agent k investigates argument $a \in \mathcal{A}_k$, let T_a be the truth value of a , and $\mathfrak{Z}_k(a)$ the score of a in the agent's opinion graph. We introduce σ , the parameter which controls the standard deviation of Gaussian distributions.

¹Each new generated argument is unique, and identified by a unique id; two agents may both generate an argument attacking a , which will translate as two new arguments attacking a . Note that there is no limit as to the number of attacks and supports an argument can have.

²Cases of seemingly conflicting arguments (eg if an agent receives both a support and an attack towards the same argument from her neighbors) do not pose any problem, since the gradual semantic is built to aggregate multiple supports and attacks.

1. The agent samples r_k from a Gaussian distribution $G_1 = \Gamma(T_a, \sigma)$. Argument a' is created. If $r_k > \mathfrak{Z}_k(a)$, the agent generates a support relation $(a', a) \in \mathcal{S}_k$ otherwise she generates an attack relation $(a', a) \in \mathcal{R}_k$.
2. T'_a is sampled from $\begin{cases} G_2 = \Gamma(T_a, \sigma) & \text{if } (a', a) \text{ is a support} \\ G_2 = \Gamma(1 - T_a, \sigma) & \text{if } (a', a) \text{ is an attack} \end{cases}$ restricted to the $[0, 1]$ interval.

This mechanism has a *corrective* effect: if the agent's opinion of the argument is higher than its true strength, the agent will likely produce an attack and therefore lower her own opinion of the argument, and vice versa. Therefore, this mechanism improves each agent's perception of the argument's strength, and its precision is controlled by σ . For this reason, we consider σ to be the agent's **experimental accuracy**.

The truth value of a supporter is similar to that of the supported argument; for attackers, the relation is reversed. Although we do not claim that this property would at any rate be verified by real-life argumentative discussions, it is a feature of our model because we wish to represent epistemic investigation and consider that agents' experiments give them access to some aspects of the world. How much the truth value of an argument influences that of its attackers and supporters is also controlled by the experimental accuracy σ .

Table 6.1: Model parameters and variables. The quantities indexed by a $*$ change at each time step t .

Symbol	Description
N	Number of agents.
p_{ER}	Probability that two agents are linked in the social network.
$\mathcal{N}$	Social network of the agents.
T_i	Truth value of the issue of the debate.
σ	Accuracy of the agent's experiments.
$\mathcal{O}_k^*$	Opinion graph of agent k .
$\mathfrak{Z}_k(a)^*$	Score of argument a in agent k 's opinion graph.
o_k^*	Opinion of agent k ; score of the the issue in agent k 's opinion graph.
T_a	Truth value of argument a .
r_k^*	Result sampled by agent k .

6.3.3 Examples

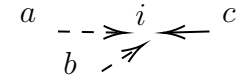
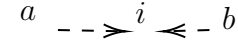
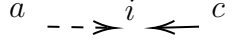
Tables 6.1 presents an overview of every parameter and variable in our model, along with their description. Here, we present an example step of our protocol.

Example 3. Let us consider $N = 3$ agents, members of the social network $\mathcal{N} = (1, 2), (1, 3)$ shown below. In the beginning, all agents' opinion graphs are composed only of the issue i . We fix $T_i = 0.8$, $\sigma = 0.5$.

$$\text{Agent 2} \xrightarrow{\mathcal{N}} \text{Agent 1} \xrightarrow{\mathcal{N}} \text{Agent 3}$$

During the first step, all agents $k \in \{1, 2, 3\}$ investigate the issue i , and they all have the same opinion $o_k^0 = 0.5$. They each sample a result r_k^0 from the same Gaussian distribution $\Gamma(0.8, 0.5)$. Table 6.2 details an example of the first step of such a run, where we indicate the sampled result, the name of the generated argument and its relation with the issue computed as stated above, its truth value, the argument that the agents learn from their neighbors and their updated opinion graph and opinion. Note that all relevant quantities are indexed for the agents k and time step $t = 0$.

Table 6.2: Example first step of a debate with three agents.

k	Opinion o_k^0	Sampled Result r_k^0	Generated Argument	Truth value	Arguments Learned	Updated Opinion Graph	Updated Opinion o_k^1
1	0.5	0.78	a , support	1.00	b, c		0.62
2	0.5	0.62	b , support	0.65	a		0.73
3	0.5	0.45	c , attack	0.3	a		0.5

At the end of this step, the three agents each have a distinct opinion graph and opinion of the issue. Agents 1 and 2's opinions are closer to the truth value of the issue $T_i = 0.8$, while agent 3's opinion has not changed, although her graph has. This shows that listening to more agents is not always beneficial for convergence to the truth, as we see in the case of agent 1.

Three questions arise naturally from the observation of our model :

- Do the opinions of the agents keep evolving indefinitely or do they stabilize in some way?
- Can the agents approach the truth, and which conditions are favorable to their epistemic success?
- Can agents influence each other in a way that makes their opinions converge?

These questions are explored in the following section.

6.4 Results

6.4.1 Stability

In theory, the agents' opinions can keep evolving forever. In practice, they vary less and less as steps go by and can reach states of "stability" where no significant variation happens for hundreds or thousands of steps. To quantify this notion, we define the **opinion stability** metric.

Definition 18 (Opinion Stability). *For an agent k , we can compute her opinion stability, the rolling standard deviation of her opinion at each time step t starting from $t = 20$:*

$$OS_k^t = \sqrt{\frac{1}{20} \sum_{j=t-19}^t (o_k^j - \mu_k^t)^2}$$

where μ_k^t is the mean of the last 20 opinions of agent k at time step t :

$$\mu_k^t = \frac{1}{20} \sum_{j=t-19}^t o_k^j$$

The opinion stability metric for the whole community at each time step t is obtained by averaging the values of S_k^t across all agents:

$$OS_t = \frac{1}{N} \sum_{k=1}^N S_k^t$$

For each agent, at each step starting at 20 we compute the standard deviation of their last 20 opinions. By averaging these values, we obtain a metric for the whole community.

6.4.1.1 Single agent

We consider simulations composed of a single agent³. Based on some qualitative analysis of debates, we consider that an agent has reached sufficient stability when her opinion stability has stayed below a threshold of 0.005 for 50 steps. Although this cannot guarantee that the opinion of the agent will not experience significant changes afterwards, this allows us to compare the speed at which different runs stabilize. We report t_s , the first step where the agent has achieved stability as defined above.

³We consider debates with a single agent to be well defined. [Mercier and Heintz \(2014\)](#) argue that scientific reasoning is argumentative even when performed by an isolated individual, as they have to anticipate the critics of their peers. In our model, the opinion of an agent evolves as they generate arguments, even if they do not communicate with others.

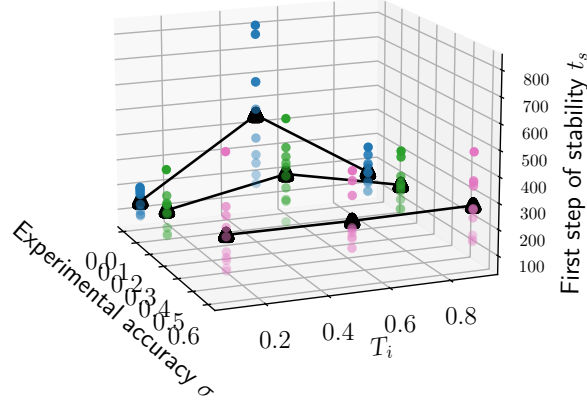


Figure 6.2: First step where stability is reached in simulations with one agent, for different values of T_i and σ . The colored points are single debate data, the black triangles are the mean for each (σ, T_i) pair.

We study the influence of the truth value of the issue T_i and experimental accuracy σ on the stabilisation speed on Figure 6.2.

The impact of T_i is clear: a value of $T_i = 0.5$ leads to significantly more variation in the agent's opinion than the extreme values at either end of the opinion space. This effect is distinct for low values of experimental accuracy σ , and imperceptible for high values of σ .⁴ Issues whose truth values are around 0.5 can be interpreted as half-true. It seems intuitive that such issues would be the hardest to settle on, compared to issues that are more clearly defined as true or false. Using this interpretation, we can state the observation below.

Observation 1 : Agents' opinions stabilize quickly when investigating issues which are close to falsity or total truth, and slowly when investigating issues whose truth values are in the middle of the opinion space.

The effect of σ can be derived from the effect of T_i : higher values mean that the agent's experimentation are less reliable, and the difference between values of T_i is blurred out. In the most extreme case, this corresponds to agents not being able to differentiate between T_i values.

⁴Additional results, which are not reported here for the sake of the visibility of the figure, confirm these effects for intermediate values of σ and T .

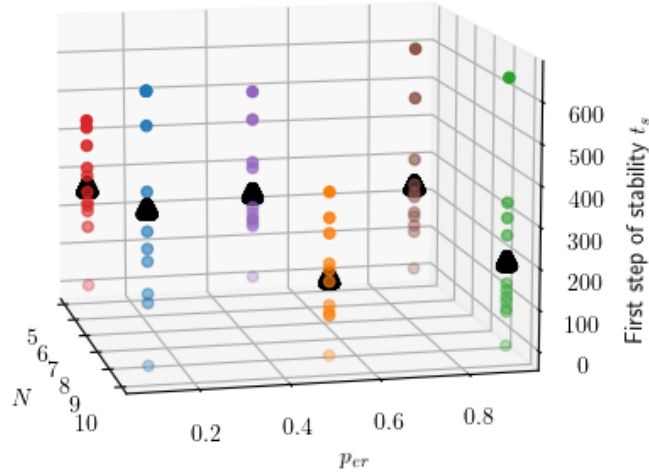


Figure 6.3: First step where stability is reached in simulations where $\sigma = 0.15$, T_i is random, and N and p_{er} vary. The colored points are single debate data, the black triangles are the mean for each (N, p_{er}) pair. A few debates did not reach stability in 600 steps, we gave them a placeholder t_s of 650.

6.4.1.2 Community of agents

Next, we want to study the impact of the number of agents N and the probability that they are connected p_{ER} on the stability. We fix $\sigma = 0.15$ and generate T_i randomly. Figure 6.3 reports the first step where stability is reached.

p_{ER} seems to be correlated with the speed of convergence, especially when $N = 10$. A high value of p_{ER} means that agents have a higher number of neighbors, and thus that their opinion graphs increase more rapidly in size. One property of issue-oriented argumentation graphs is that the probability that a new argument will have a significant impact on the issue decreases as the size of the graph increases. This can intuitively explain why well-connected communities stabilize faster.

Based on our experimental results, we decided to stop all simulations after 500 steps, as in the majority of runs we observed, the opinion of agents have reached a state of stability by then.

6.4.2 Opinion convergence

Dupuis de Tarlé et al. (2022a) show that agents' opinions converge when they exchange arguments, which hints that argumentation paired with gradual semantics could be a form of opinion-merging mechanism. Like Dupuis de Tarlé et al. (2022a), we study the convergence of agents' opinions by reporting their standard deviation, which we denote the **opinion diversity** metric.

Definition 19 (Opinion Diversity). *Let us consider a set of N agents and let o_k^t denote the opinion of agent k at time step t . The opinion diversity is the standard deviation of the opinion of agents:*

$$D^t = \sqrt{\frac{1}{N} \sum_{k=1}^N (o_k^t - \bar{o}^t)^2}$$

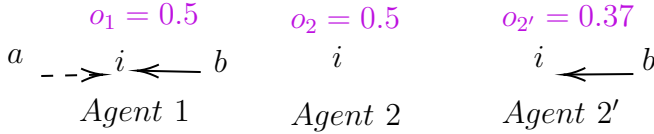
where $\bar{o}^t$ is the mean opinion of all agents at time step t :

$$\bar{o}^t = \frac{1}{N} \sum_{k=1}^N o_k^t$$

6.4.2.1 Formal discussion

Mathematically, exchanging arguments does not automatically lead to a convergence of the opinions of the agents. Consider the following example:

Example 4. *Let us consider two agents 1 and 2 whose opinion graphs, represented above, are respectively composed of two arguments and none. Their opinions are the same: $o_1 = o_2 = 0.5$. Let us consider what happens if agent 2 adds one of the arguments of agent 1 to her opinion graph: this is the situation denoted as agent 2' above. Her opinion $o_{2'} = 0.37$ is more distant to that of agent 1, even though their opinion graphs are now more similar⁵.*



A similar property was shown by [Dupuis de Tarlé et al. \(2022a\)](#) for a different semantic and class of graphs. Thus, the question of whether the agents' opinions will converge experimentally is not a trivial one.

6.4.2.2 Simulation results

Figure 6.4 reports the opinion diversity of the community when we vary p_{ER} , the probability that two agents are connected in the social network. We observe a surprising threshold phenomenon: the opinion diversity *increases* with p_{ER} as long as $p_{ER} \leq 0.5$, then *decreases* when $p_{ER} \geq 0.5$ to reach 0 in the case of a complete network.

Observation 1 : Communities become more diverse when p_{ER} increases below 0.5 and less diverse when p_{ER} increases above 0.5.

⁵The notion of similarity between graphs is not developed here, but consider for example a graph-edit distance ([Sanfeliu and Fu, 1983](#)).

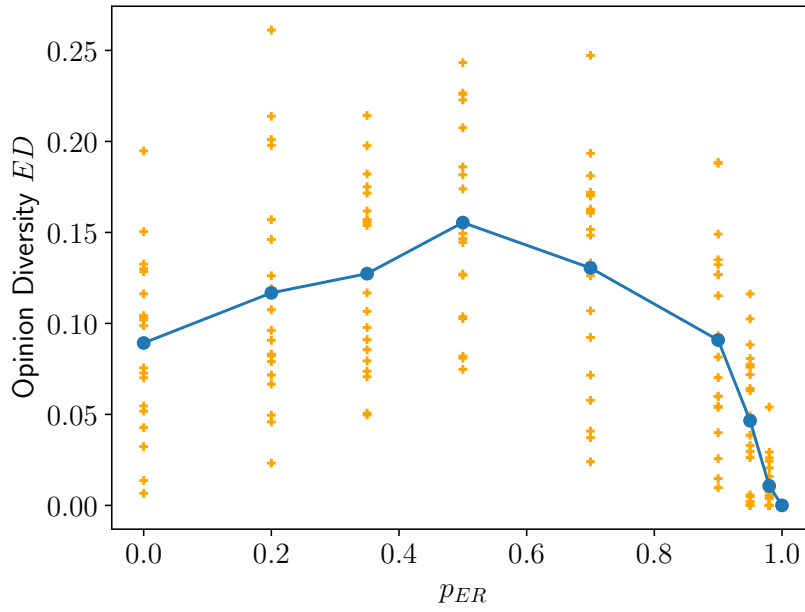


Figure 6.4: Opinion diversity. $N = 10$, $\sigma = 0.15$, T_i is random, p_{ER} varies. The orange points are individual simulation results, blue points are averaged for each p_{ER} .

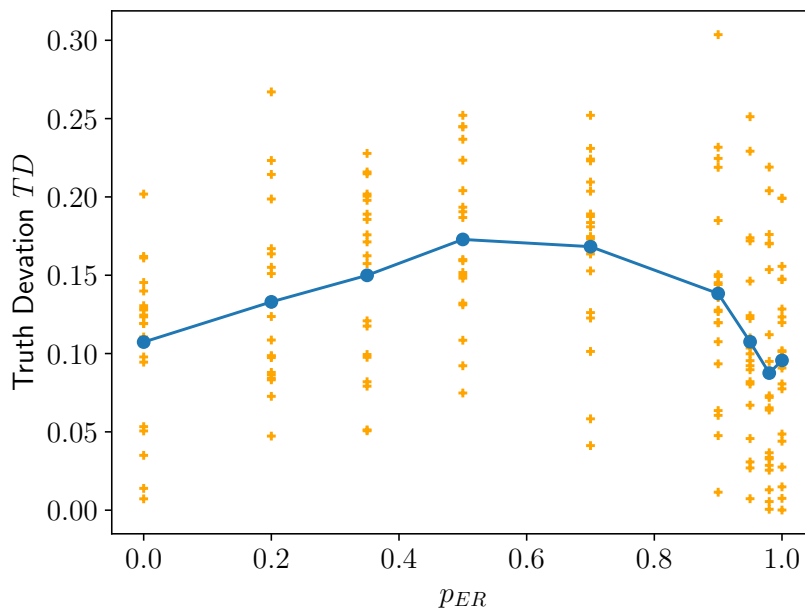


Figure 6.5: Truth deviation. $N = 10$, $\sigma = 0.15$, T_i is random, p_{ER} varies. The orange points are individual simulation results, blue points are averaged for each p_{ER} .

These results are surprising. When p_{ER} is low, agents interact with a limited number of other agents, but there is no obvious reason that this should lead to have more diverse opinions as a whole: in the classical bounded confidence models for example, interaction between agents always leads to an averaging of their opinion and thus, to a decrease of global diversity. Our model's dynamics are more complex, as we have formally shown that communication may lead to divergence, and as the opinion of our agents evolve even when they do not interact with each other. Further analysis is warranted to understand these effects.

6.4.3 Epistemic success

The epistemic success of agents refers to the degree to which they are able to arrive at accurate opinions. In our case, this accuracy is measured as the distance between an agent's opinion and the truth value T_i . We use the **truth deviation** metric introduced by [Hegselmann et al. \(2006\)](#) to measure the epistemic success of our community.

Definition 20 (Truth Deviation). *Let us consider a set of N agents and let o_k^t denote the opinion of agent k at time step t . The truth deviation is:*

$$TD^t = \sqrt{\frac{1}{N} \sum_{k=1}^N (o_k^t - T_i)^2}$$

with T_i the truth value of the issue of the debate.

6.4.3.1 Experimental accuracy

We find that σ is correlated with the truth deviation: the smaller it is, the better agents are at approximating the truth (see Figure 6.6). This result is not unexpected given the role of σ as the experiment accuracy of the agents. It validates our model as one where the success of truth-seeking agents can be controlled by an accuracy parameter.

6.4.3.2 Probability of connection

When we vary p_{ER} (the probability that two agents are connected) in a community of $N = 10$ agents, we see in Figure 6.5 that the truth deviation *increases* for $p_{ER} \leq 0.5$ then *decreases* for $p_{ER} \geq 0.5$. We observe the same threshold of $p_{ER} = 0.5$ as in the opinion diversity metric. This hints of a relation between the diversity of the opinions and the truth deviation. To explore this link, we plot in Figure 6.7 the opinion diversity of each of the debates with varying p_{ER} and the truth deviation of

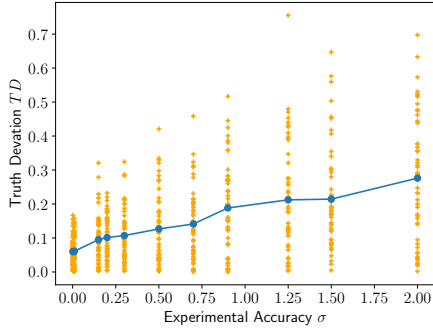


Figure 6.6: Truth deviation. $N = 1$, σ varies, T_i is random. The orange points are individual simulation results, blue points are averaged for each p_{ER} .

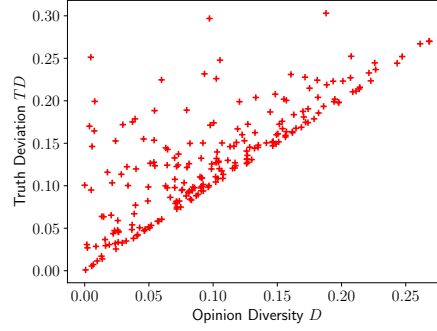


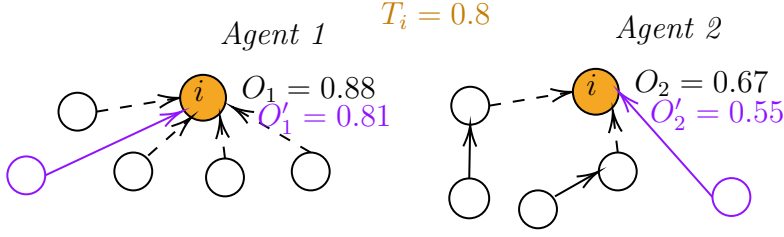
Figure 6.7: Truth deviation against opinion diversity. $N = 10$, $\sigma = 0.15$, T_i is random, p_{ER} varies.

the community, and we see clearly that higher diversity is associated with a worse epistemic success.

Observation 2: Diverse communities are less epistemically successful.

Remarkably, this result presents similarities with the unintuitive Zollman paradox which states that more communication is sometimes detrimental to the epistemic success of communities (Zollman, 2007). In Zollman’s model, however, less communication is correlated with more diversity, which gives communities a boost in epistemic success. In our model, the relationship between connectivity and diversity is more complex, and diversity worsens the outcome of agents. Zollman’s model is composed of Bayesian agents, who exchange directly the result of their investigations, whereas our agents transform their investigations into arguments. Given how we modelled argument generation, an argument improves the opinion of the agent who generated it, because it corrects her current opinion. However, if it is shared with another agent who holds a different opinion, it can mislead her: this could explain the harmful effect of diversity.

Example 5. Consider the following situation: the truth value of the issue is $T_i = 0.8$, two agents are equipped with opinion graphs (without the purple argument), thus have opinions $O_1 = 0.88$ and $O_2 = 0.67$. Now, let’s consider what happens if agent 1 investigates the issue, and collects a sample result $r_1 = 0.8$. Because $O_1 > r_1$, this sample result is transformed into an attacking argument by agent 1. After generating and adding this new argument (in purple), the opinion of agent 1 is now 0.81, closer to the truth value of the issue. However, if agent 1 shares this new argument with agent 2, then the opinion of agent 2 decreases as well, to 0.55: it is now further away from the truth value of the issue. The difference between their opinions made the exchange of arguments counter-productive.



6.5 Discussion and Conclusion

We present a novel modelling of an epistemic community using abstract argumentation as the representation of information. The analysis of our model presents us with three stylized facts.

1. Agents' opinions stabilize quickly when investigating issues which are close to falsity or total truth, and slowly when investigating issues whose truth values are in the middle of the opinion space.
2. Communities become more diverse when p_{ER} increases below 0.5 and less diverse when p_{ER} increases above 0.5.
3. Diverse communities are less epistemically successful.

Observation 1 is an emergent result which hints that middle ground answers are harder to settle on than clear-cut answers. This intuitive result indicates that our model generates interesting dynamics reflective of real-world epistemic processes.

Observation 2 is of particular interest. One limit of classical opinion diffusion models is that they cannot account for phenomena of polarisation as an increase in the distance between people's opinions which have been reported by real-life experiments (Bramson et al., 2017a). Observation 2 suggests that our model may account both for an increase in the diversity of opinions when agents interact more, as well as the formation of a consensus in tight-knit communities.

Observation 3 enriches the discussion about the counter-intuitive effects of communication in epistemic communities, opened by Zollman's paradox (Zollman, 2007). Our results hint at a more complex link between the rate of communication between agents and the diversity of their opinions, and our observation is in direct contrast with agent-based models from the literature showing that opinion diversity benefits truth-seeking communities (Hong and Page, 2004; Gabriel and O'Connor, 2022). The results of our model also contrast with that of Borg et al. (2017a), who find that communication increases the epistemic success of their agents. As their model is also argumentative, this hints that differences in the protocols are decisive in the final outcome. One direction for our future work will be to explore how our

model can inform the role of the mode and quality of communication on the effect of diversity in epistemic communities, as explored in ([Sulik et al., 2022](#)).

CHAPTER 7

An Agent-Based Model of Myside Bias in Scientific Debates

Recent insights in cognitive science suggest that scientific reasoning is influenced by *myside bias*. This bias manifests as a tendency to prioritize the search and generation of arguments that support one’s views rather than arguments that undermine them. Additionally, individuals tend to apply more critical scrutiny to opposing stances than to their own. Although myside bias may pull individual scientists away from the truth, its effects on communities of reasoners remain unclear.

In this chapter, we present an agent-based model for studying the impact of myside bias on the argumentative dynamics in scientific communities. This work was recently published in the *Journal of Artificial Societies and Social Simulations* in [Dupuis de Tarlé et al. \(2024\)](#).

The aim of our model is two-fold: first, to study the argumentative dynamics generated by myside bias, and second, to explore which mechanisms may act as a mitigating factor against its pernicious effects. Our results indicate that biased communities are epistemically less successful than non-biased ones, and that they also tend to be less polarized than non-biased ones. Moreover, we find that two socio-epistemic mechanisms help communities to mitigate the effect of the bias: the presence of a common filter on weak arguments, which can be interpreted as shared beliefs, and an equal number of agents in favor of each alternative at the start of the scientific debate.

7.1 Introduction

The process of scientific inquiry has long intrigued social scientists and philosophers of science. How can individual scientists, driven by various incentives, jointly form a community that effectively and efficiently achieves its epistemic goals? Inspired by formal approaches in economics, the late '90ies gave rise to various analytical models examining the relationship between individual and group rationality (e.g. [\(Kitcher, 1990; Zamora Bonilla, 1999; Brock and Durlauf, 1999\)](#)). Soon after, with the introduction of computational methods, more complex scenarios have been investigated. In particular, simulations in the form of agent-based models (ABMs) were introduced to study how local interactions among individual scientists bring about emergent phenomena characterizing collective inquiry (e.g. [\(Gilbert, 1997; Hegselmann et al., 2006; Zollman, 2007\)](#)).

While various ABMs were developed to study the process of scientific interaction, the argumentative aspect of scientific deliberation has only recently gained attention in this body of literature. This is surprising given that argumentation has traditionally been considered of central importance for the progress of science (Popper, 1962; Kuhn, 2000; Pera, 1994; Rescher, 2007), and the idea that knowledge production relies on a continuous scientific debate, in which arguments are criticized, rejected or defended, is widely recognized as the cornerstone of scientific objectivity (Douglas, 2009). In fact, although individual scientists may suffer from various biases, the continuous critical exchange is meant to keep these biases in check (Longino, 2002). At the same time, we lack a precise understanding of mechanisms that allow a scientific community to achieve its goals despite the pervasive presence of biases (Zollman, 2011).

This chapter aims to examine one such bias which has gained significant attention in cognitive science: what Mercier, Sperber and their colleagues define as *myside bias* (Mercier, 2016). They argue that reasoning is inherently argumentative, and rather than being focused on tracking truth, it fulfills two argumentative functions: to convince an audience and to evaluate the arguments of other agents (Mercier and Sperber, 2017; Mercier, 2017). In their view, myside bias concerns the first function, where reasoning operates in a biased way by predominantly producing arguments in favor of held views rather than seeking out defeating arguments. We stick here to the conceptualisation by Mercier, Sperber and colleagues (Mercier and Heintz, 2014; Mercier and Sperber, 2011; Mercier et al., 2016) that only consider myside bias as *production* bias, and differentiate from other uses of the notion, where either production, evaluation or both are concerned (Stanovich et al., 2013; Mercier, 2016; Wolfe and Britt, 2008; Baron, 1995).¹ The characterization of myside bias in terms of the production of reasons renders myside bias essentially part of argumentative reasoning. As such, it affects various reasoning contexts, including scientific reasoning (Mercier and Heintz, 2014).

While myside bias may pull scientists away from the truth, the effect is most disastrous when scientists reason isolated from a social context. As soon as agents exchange arguments, argument evaluation takes place, where reasoning is "quite good at telling apart good from bad arguments". Therefore, Mercier and Heintz (2014) conjecture that the detrimental effects may be absorbed in the social context of a scientific community in which *pro* and *con* arguments are exchanged and critically evaluated.

Moreover, potential adverse effects of myside bias can be further mitigated when scientists share evaluative standards for what counts as a good argument (Mercier and Heintz, 2014). Shared evaluative standards consist of shared "beliefs about what a good argument looks like" (ibid, p. 519), which enable agents of a community to weed out weak arguments. As Mercier and Heintz (2014) argue, different scientific domains come with more or less such shared standards. While mathematics comes

¹The term myside bias was coined in an unpublished manuscript by Perkins from 1986, eventually published in Perkins (2019).

with a rather strict regiment of what counts as a proof, we see more variety and dynamics in other disciplines such as psychology, where in response to the replication crisis new standards obtain increasing significance (such as preregistered studies, etc.). Mercier and Heintz look at these shared beliefs as a possible tool for scientific community to contrast the effect of bias, insofar as they help weeding out poor arguments.

This view contrasts sharply with a more traditional take on confirmation bias as generally epistemically pernicious. At the same time, it coheres with recent proposals that confirmation bias, while harmful for an isolated reasoner, may not be detrimental to group inquiry (Smart, 2018; Peters, 2020). Therefore, the current state of the art raises the question: How exactly does myside bias impact collective inquiry?

We aim to address this issue. In particular, we will examine two closely related questions:

1. What kind of effect (if any) does the bias create in a scientific community that exchanges arguments?
2. Do shared beliefs in the form of shared evaluative standards among scientists help to mitigate the bias's detrimental effects?

Answering these questions is crucial to better understand the impact of myside bias on scientific debates and collective inquiry in general.

To investigate these issues, we develop an ABM² based on the abstract argumentation framework (Dung, 1995). Abstract argumentation has previously been used for the formal modeling of scientific debates (Šešelja and Straßer, 2013) as well as in agent-based modeling of scientific inquiry (Borg et al., 2017a,b). At the same time, applications of abstract argumentation to the study of biases, and the phenomenon of myside bias in particular, have been lacking. Modeling a scientific debate in terms of abstract argumentation allows us to represent myside bias in line with Mercier and Sperber's view: as a cognitive mechanism that affects the production of arguments. In addition, it also allows us to test whether the presence of shared evaluative standards (Mercier and Heintz, 2014) acts as a mitigating factor against the insidious effects of the bias.

The chapter is structured as follows. Section 2 provides the necessary background by introducing myside-bias in more detail and our choice of argumentation framework underpinning our model. In Section 3, we present our model and in

²A description of the model following the ODD protocol and the source code can be found on the CoMSES database: <https://www.comses.net/codebase-release/49d2dbe8-7f8e-43f2-a27f-87afdac7f5f7/>

Section 4, the central results of the simulations. In Section 5, we compare our contribution with other formal models that have addressed similar questions. Section 6 concludes the chapter.

7.2 Myside Bias and Scientific Argumentation

Myside bias affects the production of arguments of scientists, and our research questions **Q1** and **Q2** concern its effects on scientific inquiry. Given the complexity underlying scientific argumentative exchange, we propose to investigate these questions with an agent-based model that represents the argument exchange through *abstract argumentation* (Dung, 1995). This section motivates our choice.

7.2.1 Myside bias as a production bias

Mercier, Sperber and colleagues consider myside bias as an individual bias that affects the way arguments are produced. Yet, because of the extreme complexity of the scientific debate, the way such a bias impacts the epistemic performance of a scientific community cannot easily be derived in the absence of a clear formalization of the way agents interact. To see this consider a concrete example.

Example 6. *In the first half of the twentieth century, the earth scientist Alfred Wegener proposed the continental drift hypothesis, which postulated that continents used to be connected and had subsequently drifted apart to their current locations. Wegener’s theory (‘Drift’) conflicted with two other theories endorsed by earth scientists at the time: ‘Contractionism’, according to which the phenomena on the surface of the Earth are explained by the Earth’s gradual cooling and contracting, and ‘Permanentism’, which postulated that no major lateral shifts occurred on the Earth’s surface (Šešelja and Weber, 2012; Šešelja and Straßer, 2013). Wegener’s theory offered a unifying explanation of various phenomena – from the jigsaw-like fit of the continental coastlines to the similarity between flora and fauna on the opposite sides of the ocean. However, it suffered from a major conceptual issue: the lack of a plausible mechanism of the Drift, explaining how continents could move through (what appeared to be) the hard ocean floor.*

Suppose now that scientists involved in such a debate were affected by myside bias. How could have the myside bias influenced the debate? From the individual perspective, the bias make agents more likely to produce arguments in favor of their own theory than arguments critical of it, and more likely to make arguments critical of rival theories than arguments that would support them (Mercier, 2016; Mercier and Sperber, 2017). In the drift debate, this may have resulted in ‘biased drifters’ who were more likely to produce arguments in favor of the explanatory power of Drift than those elaborating on the lack of a plausible mechanism of drifting continents.

In contrast, ‘biased contractionists’ may have been more likely to produce the latter kind of arguments than those concerning problems within their own theory (such as the presence of radioactive materials in the Earth, which indicated the Earth could be heating up rather than cooling down).

However, our intuitions let us down when it comes to envisioning how the bias affected the collective debate. Supposing agents were biased, did the bias actually delay the production of critical arguments? Or perhaps its presence had little to no influence given that it was equally distributed across the proponents of all the competing theories? Did the fact that ‘drifters’ were a minority in the community modified the collective impact of bias?

Abstract argumentation and ABMs are the ideal tools to model how myside bias affects a debate. Abstract argumentation has previously been used for the formal modeling of scientific debates ([Šešelja and Straßer, 2013](#)) as well as in agent-based modeling of scientific inquiry ([Borg et al., 2017a,b](#)). It provides a precise and schematic representation of the evolution of a debate. Agent-based modeling enables the study of the collective outcome starting from the individual interactions.

Although argumentation and various forms of biased reasoning have been studied through ABMs, no ABM, to our knowledge, has focused on the combination of argumentation and myside bias intended as a production bias. On the one hand, ABMs studying argumentative dynamics have mainly focused on the argumentative exchange as free from the intrusion of cognitive biases (e.g. ([Mäs and Flache, 2013](#); [Borg et al., 2017a](#); [Kopecky, 2022](#); [Taillandier et al., 2021](#))). On the other hand, ABMs studying the impact of confirmation bias on the scientific inquiry have tackled this issue independently of argumentation ([Baccini and Hartmann, 2022](#); [Gabriel and O’Connor, 2022](#); [Baccini et al., 2023](#)). Confirmation bias is here considered as impacting the evaluation of one’s beliefs (e.g. by discounting opposing evidence) rather than affecting the production of one’s arguments. Similarly, ABMs studying how biases affect argumentative exchange have focused on evaluative aspects of biased reasoning, and on broader contexts of deliberation, not aimed at inquiry and scientific goals in particular ([Banisch and Shamon, 2021](#); [Singer et al., 2018](#); [Dupuis de Tarlé et al., 2022a](#); [Proietti and Chiarella, 2023](#)).

Now, before proceeding to the presentation of our model (in [Section 7.3.1](#)), we introduce the framework of abstract argumentation. We do so by providing some useful notions and an illustrative example of how abstract argumentation may model the debate of [Example 6](#).

7.2.2 Abstract argumentation

To simulate a biased production of arguments in a community of deliberating agents, we first need a suitable modeling framework. A particularly suitable framework for

the formal study of argumentation is *abstract argumentation*. Originally developed in symbolic AI (Dung, 1995), abstract argumentation has been increasingly used for agent-based modeling of deliberating communities, as we developed in Chapters 3 and 4. Here, we remind the notions of argumentation frameworks, path, attackers and defenders which we introduced in Chapter 2.

Formally, an abstract argumentation framework (AF) is a directed graph in which the nodes are abstract representations of arguments and edges represent argumentative attacks. In this way, AFs can be used to model a stance of a rational reasoner in a debate: by assigning to each agent a specific argumentation graph we can represent a set of arguments and attacks between them, forming the agent's argumentation map. Moreover, we can define the rules specifying which arguments are acceptable, or which arguments should be rejected.

Definition 21 (Argumentation framework). An **argumentation framework** (AF) is a pair $\langle \mathcal{A}, \mathcal{R} \rangle$ where $\mathcal{A}$ is a finite and non-empty set of (abstract) arguments, and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation on $\mathcal{A}$ called the attack relation. For two arguments $a, b \in \mathcal{A}$, we say that a attacks b in case $(a, b) \in \mathcal{R}$.

Definition 22 (Path). Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $a_1, a_2, \dots, a_n \in \mathcal{A}$ be n arguments, with $n \in \mathbb{N}$. We say that $(a_1, a_2, \dots, a_n)$ is a **path** from a_1 to a_n if for every $i \in \llbracket 1; n \rrbracket$, $(a_i, a_{i+1}) \in \mathcal{R}$.

Definition 23 (Attacker and Defender). Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $a, b \in \mathcal{A}$ be two arguments. a is a **defender** (resp. **attacker**) of b if it is situated at the beginning of an even-length (resp. odd-length) path towards b .

As we said in Chapter 2, Dung defined different ways in which sets of arguments can be deemed acceptable in a given argumentation framework. These sets intuitively characterize rational positions in a debate. In this chapter we use the grounded set due to its uniqueness.³

Definition 24 (Grounded set). Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF. An argument $a \in \mathcal{A}$ is **acceptable with respect to a set of arguments** E in case for every b for which $(b, a) \in \mathcal{R}$, there is a $c \in E$ for which $(c, b) \in \mathcal{R}$. The **grounded set** G for AF is iteratively defined as follows: $G = \bigcup_{i=0}^{\infty} G_i$ where G_0 is the set of all arguments which have no attackers in $\mathcal{A}$ and G_{i+1} is the set of all arguments that are acceptable with respect to G_i . We call an argument **acceptable** if it is part of the grounded set.

The following example illustrates how abstract argumentation models a scientific discussion.

Example 7. Imagine how a stylized discussion between Alfred Wegener, proponent of 'Drift', and Ernest Ingersoll, proponent of 'Permanentism', may have played out

³It is worth noticing that the other semantics defined in (Dung, 1995) coincide with the grounded semantics in the AFs underlying our ABM, which are trees by construction (see Sec.7.3.1).

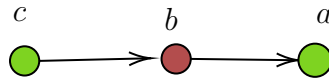


Figure 7.1: Representation of the debate of Example 7 as an argumentation graph, where the arguments are represented as nodes and the attack relations as directed edges. a , c in light green are acceptable, and b in dark red is not.

(see Example 6). Wegener could have stated argument a : ‘since the continents look like a jigsaw puzzle, they must have drifted away from each other’. Ingersoll could have pointed out that ‘although they may look as a jigsaw puzzle from a distance, their borders do not really match’ (argument b). Finally, Wegener could have rebutted Ingersoll’s argument by suggesting that ‘the differences must have been caused by erosion’, which would constitute argument c . The argumentation graph capturing this debate is illustrated in Figure 7.1. While b attacks a , c attacks b and thus defends a . Given this stage of the debate, we can say that a set of arguments $S = a, c$ is grounded and acceptable.

As the example illustrates, the argumentation map of a reasoner may grow indefinitely, with new counterarguments being added to the graph.

7.3 The Model

In this section, we introduce our model. We start by describing how agents evaluate a debate and update their preferences. This will help us to define the general setting of our simulations. Subsequently, we lay out our model’s protocol, starting with the baseline case, and only after introducing the modelling of myside bias. Finally, we detail the metrics we use to measure a community’s epistemic success and polarisation.

7.3.1 Modelling opposing views with argumentation frameworks

Our model represents a scientific community whose members try to decide between two competing alternative *research programs*. In the context of science, they could be, for example, competing scientific theories or explanations. Informally speaking, a research program is constituted by a certain number of claims.

Definition 25 (Research Program). A **research program** (*ResProg*) is a set of arguments of size N_{NA} (≥ 1).

We call the arguments included in a research program *central arguments*. Central arguments provide the central theoretical tenets of a research program. They give

the argumentative support to the core hypothesis of the program. For example, in the continental drift research program by Wegener, some of the central arguments were: (a) Jigsaw fit of the continental coastlines, indicating that the continents were once connected; (b) Similarity of flora and fauna on the opposite sides of the ocean, indicating that the biological systems on different continents were once united; etc. Both research programs contain the same number of distinct central arguments.

Agents engage in a debate by discussing the central arguments of two opposing research programs, $ResProg_1$ and $ResProg_2$. For this, they produce additional (dialectical) arguments that are not contained in research programs and which are used to defeat and/or defend central arguments. For example, the similarity of flora and fauna can be attacked by adaptation to similar climate and geological circumstances.

Each agent i has her own view of the debate at time t . The view of an agent determines what agent i thinks is the state of the art regarding the discussion between $ResProg_1$ and $ResProg_2$ at time t .

Definition 26 (View of Debate). *The view of the debate V_i^t of agent i at step t is an argumentation framework $\langle \mathcal{A}_i^t, \mathcal{R}_i^t \rangle$ such that for any t, i : $ResProg_1, ResProg_2 \subset \mathcal{A}_i^t$.*

As such the view of an agent always contains the central arguments of each research program, and may then include new arguments that are relevant to the debate. Arguments may be connected by attack relations. An attack relation from an argument x to an argument y represents an explicit argumentative move that challenges argument y . Such an argumentative move may be obtained through empirically refuting one important step of the attacked argument, by pointing out a methodological mistake and so on. As a simplification, no attack relation is present among the central arguments given their supportive character.⁴ Because of how views are built (see Section [Baseline Protocol](#)), they have two special features. First, each central argument of a research program is the root of an in-tree, that is a tree where all edges point towards the root. Second, each argument $a \in \mathcal{A}_i$ belongs to one and only one of the in-trees that can be found in the framework, that is every non-central argument can be linked by a path with one and only one of the central arguments. Figure 7.2 presents an example of an in-tree with its own root, and Figure 7.3 presents an example of a view. While the view of every agent contains the central arguments of each research program, different agents can have different views.

Based on her own view of the debate, an agent assigns an individual score to each research program, by applying the acceptability semantics. The score of a research program is equal to the number of its central arguments that are acceptable in her view.

⁴Although central arguments do not attack each other, they can be incompatible ((Šešelja and Straßer, 2013)), insofar as they reach opposite conclusions. Because agents only support one research program at a time, they never end up supporting incompatible arguments.

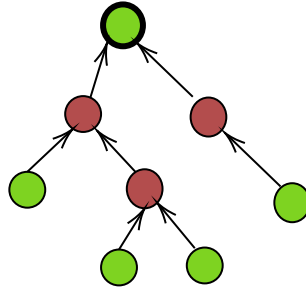


Figure 7.2: Example of an in-tree. The acceptable arguments are colored in light green, the others in dark red. The central argument is outlined in bold.

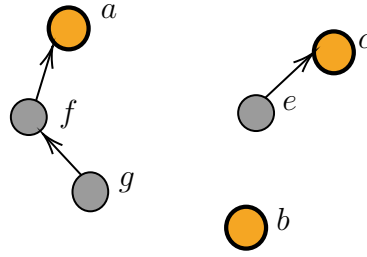


Figure 7.3: Schematic representation of a possible view of an agent composed of three in-trees with central arguments a , b and c (in orange and bold).

Definition 27 (Individual Score of a Research Program). *Let i be an agent with view V_i^t at time t . Then, the score agent i assigns to $ResProg_j$ at time t is equal to*

$$score_i^t(ResProg_j) = |\{a \in ResProg_j | a \text{ is acceptable in } V_i^t\}|.$$

Whenever one of the two research program has a higher score than the other one, we say that an agent prefers it. When agents have different views, they may assign different scores to the same research programs, and, consequently, have different preferences for research programs.

Finally, we introduce a function S , which assigns a dialectic strength to each argument, in short **strength**. Whenever an agent looks for a counterargument to a , the probability with which such an attack is generated depends on $S(a)$. The variability in argument strength accounts for the intuitive observation that some arguments are more convincing and harder to attack than others. We assume that the central arguments of a research program have the same strength. This modeling choice allows us to decide which (if any) of the two research programs is objectively stronger and, consequently, objectively preferable. A more robust research program represents a research program that is based on stronger evidence or provides better explanations than the rival one.

We call S_1 and S_2 , respectively, the values for the strength of the central arguments in $ResProg_1$ and in $ResProg_2$, i.e. if a is a central argument of $ResProg_1$, $S(a) = S_1$. Section [Baseline Protocol](#) explains how the value $S(a)$ is computed if a

is not a central argument. We refer from now on to S_i interchangeably as either ‘the strength of $ResProg_i$ ’ or as ‘the strength of the central arguments of $ResProg_i$ ’. The strengths of the two research programs S_1, S_2 are treated as simulation parameters.⁵

Our model simulates an ongoing debate among scientists about the merits of two research programs $ResProg_1$ and $ResProg_2$. It does so through a succession of T steps in which agents generate new arguments and argumentative attacks and update their views and preferences for research programs. One could think of arguments in terms of the argumentative content of research papers published by scientists within a certain debate. We begin by outlining the baseline protocol, and then proceed to the following two subsections, where we describe the implementation of myside bias and a specific mechanism aimed at mitigating its harmful impact.

7.3.2 Baseline protocol

Initialization Each agent i is initialized with a view whose argument set contains only the central arguments of both research programs and no attack relation, i.e. $V_i^0 = \langle \mathcal{A}_i^0, \mathcal{R}_i^0 \rangle$ such that $\mathcal{A}_i^0 = ResProg_1 \cup ResProg_2$ and $\mathcal{R}_i^0 = \emptyset$ (see Figure 7.4 for an example). Since all agents have access to the same set of arguments, their views are identical at the start of the simulation and cannot be used to determine their preferences. The preference of each agent at step 0 is defined by parameter n_1^0 (called initial support), which specifies the number of agents who prefer $ResProg_1$. All of the remaining agents prefer $ResProg_2$.

Protocol Each step has three phases.

Phase 1: “Attack generation”.

- Each agent i randomly chooses some argument a among those present in her set of arguments $\mathcal{A}_i^t$.
- Each agent i tries to generate a counterargument to a , that is, a new attacker b of a . The success rate of generating the counterargument is $p_{attack} = 1 - S(a)$ where $S(a)$ is the strength of a . If successful, the new argument b obtains a strength $S(b)$ drawn from a normal distribution centered on $1 - S(a)$ with standard deviation $StDev$. The argument generation probability p_{attack} is based on the idea that attacking weak arguments is easier than attacking strong ones. The strength generation is based on the idea that attackers of strong arguments will be on average weaker, and vice versa.⁶

⁵Note that the strength of each research program is not accessible to the agents and thus does not directly impact their preference. It does impact the agents’ preference indirectly via the impact it has on the process of argument generation.

⁶Agents produce arguments simultaneously. To keep things simple and focused on the effects of myside bias, we avoid modeling priming effects.

Phase 2: “Filtering the arguments”.

- For each generated argument b , the community decides whether it should be *considered*. The default assumption is that the community considers all the generated arguments, unless it employs specific mechanisms for their filtering (we will introduce one such mechanism in Section 7.3.4).

Phase 3: “Update of views and preferences”.

- The agent i who produced argument b always adds argument b and the correspondent attack relation (b, a) to her view (regardless of whether the community has filtered that argument). If the community decides the argument b should be considered, each other agent may include it or not in their view. Every agent j who does not know argument a , that is $a \notin \mathcal{A}_j$, does not add argument b nor relation (b, a) to her view. Finally, an agent k who knows a but did not produce b adds the argument b and the relation (b, a) to her view with probability p_{see} .
- Each agent i computes the score for each research program based on her view (Definition 27) and updates her preference accordingly. She prefers the *ResProg* with the highest score, and does not change her opinion if the two programs are equal.

Through this protocol, the views of the agents are populated. In the beginning the only arguments the agents know are the central arguments of each research program, and a new argument can only attack an argument that is already present. As a consequence, each central argument of a research program is the root of an in-tree, that is a tree where all edges point towards the root, and no cycle can be formed (similarly to (Dupuis de Tarlé et al., 2022a; Proietti and Chiarella, 2023)). This reflects the unfolding of a scientific debate, in which the new arguments produced are usually unlikely to be attacked by arguments that are already present. Notably, if the probability of including an argument for the agents in the community is lower than 1 ($p_{see} < 1$), the agents will have distinct views which will not contain every argument that has been produced.

Our agents evaluate the debate impartially, insofar as they always support the research program with the highest score. The score of a program corresponds to the number of its central arguments that are acceptable. As such, it indicates the strength of a program given the present state of the debate: the program with the higher score is the one that appears as the stronger one.

Example 8. *Let us illustrate this protocol with the first step of a debate concerning two research programs of strength $S_1 = 0.5$ and $S_2 = 0.8$. Each research program contains two central arguments, as shown in Figure 7.4: arguments a and b for $ResProg_1$ and c and d for $ResProg_2$. For the sake of simplicity, we consider the*

view of only a single agent $Agent_1$ who begins the simulation with a preference for $ResProg_1$. Her view at time 0 (V_1^0) is shown in Figure 7.4, .

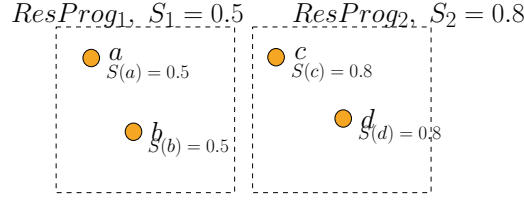


Figure 7.4: View V_1^0 of the debate for agent 1 at the start of the simulation.

First, $Agent_1$ randomly selects an argument from her view (V_1^0), in this case a . The agent tries to generate an attacking argument with a probability $p_{attack} = 1 - S(a) = 0.5$. The agent succeeds: a new argument e that attacks a is created. The strength of e is sampled from a normal distribution of mean $1 - S(a) = 0.5$. We obtain a strength of 0.43 ($S(e) = 0.43$). The view at time 1 (V_1^1) shown in Figure 7.5 is the result of updating the initial view by adding the new argument e and an attack between e and a .

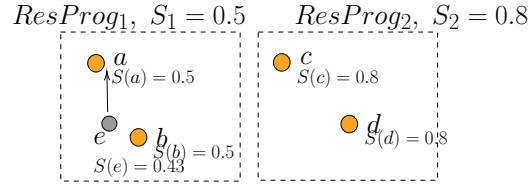


Figure 7.5: View V_1^1 of the debate.

Then, $Agent_1$ computes the score of each research program by counting the number of acceptable central arguments in her view. Here, arguments b , c and d are not attacked and are thus acceptable, while a is attacked but not defended which makes it not acceptable. The scores of each research program for agent 1 are $score(ResProg_1) = 1$ and $score(ResProg_2) = 2$. Since $ResProg_2$ has a higher score than $ResProg_1$, $Agent_1$ changes her preference to $ResProg_2$ for the next step.

7.3.3 MySide Bias in the production of arguments

As mentioned above, myside bias is modelled as a factor that makes it easier to *produce* arguments in favor of one's standpoint. In our protocol, the bias increases the probability of generating a counterargument in case the counterargument is *in favor* of the agent's preferred research program. An argument favors one's preferred research program if it either defends the program or attacks the opposing one (recall Definition 23).

Definition 28 (Against - In favour of). An argument is said to be **in favour** of a research program $ResProg_i$ and **against** the other $ResProg_j$ if it is a defender of one of the central arguments of $ResProg_i$, or an attacker of one of the central arguments of $ResProg_j$.

We model a biased community by changing the process of argument generation as follows. First, each agent is equipped with a bias parameter $bias \in [0, 1]$. If an agent is investigating an argument a *in favor* of her preferred research program, then

$$p_{attack} = \begin{cases} 1 - S(a) - bias & \text{if } 1 - S(a) - bias \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad (7.1)$$

If she is investigating an argument that is *against* her preferred research program then

$$p_{attack} = \begin{cases} 1 - S(a) + bias & \text{if } 1 - S(a) + bias \leq 1, \\ 1 & \text{otherwise.} \end{cases} \quad (7.2)$$

The bias parameter changes the probability with which an agent finds a counter-argument against a certain argument, based on the agent's preference. The higher the parameter is, the more biased the agents are. When $bias = 0$ for all agents, we are in the baseline case of an unbiased community.

Example 9. *To illustrate the difference induced by the bias, consider the first step of the same debate of Example 8, where Agent₁ is now biased with $bias = 0.3$. The view of Agent₁ is the same at the beginning (Figure 7.4), and she begins the simulation with a preference for ResProg₁.*

Agent₁ randomly selects an argument from V_1^0 , in this case a . Attacking a would be against the agent's preferred research program, ResProg₁. The agent tries to generate an attacking argument with a probability $p_{attack} = \max(1 - S(a) - bias, 0) = 0.2$. The agent is more likely to fail than in the non-biased case. Let us assume that she fails to generate an attack towards a . Her view does not change in the next step.

Agent₁ computes the score of each research program by counting the number of acceptable central arguments in her view. Here, arguments a , b , c and d are not attacked and thus acceptable. The scores of the two research programs for Agent₁ are the same: $score(ResProg_1) = 2$ and $score(ResProg_2) = 2$. Since $score(ResProg_1) = score(ResProg_2)$, agent 1 keeps preferring ResProg₁ in the next round.

This example highlights how the bias makes it harder to generate arguments disfavoring one's preferred research program, even though the arguments one is trying to attack may be weak. As such, the bias may comfort the agents in their opinions, and be detrimental to their assessment of the best research program. In this case, it is as though her bias prevents Agent₁ from seeing the faults of her preferred point of view.

Recall that we assume biased agents to evaluate the debate and choose which program to support using the same procedure as non-biased agents. This is a crucial aspect of the account of myside bias by [Mercier and Heintz \(2014\)](#), who argue that

myside bias does not affect the evaluation of arguments, but only their production. Our agents are biased when they generate arguments, not when they evaluate the research programs.

7.3.4 Shared beliefs as shared standards of evaluation

Mercier and Heintz (2014) suggest that the presence of shared beliefs may help a community to perform better and exhibit reduced susceptibility to the harmful effects of myside bias. The idea is that the production of arguments is improved if agents can anticipate counter-arguments from other agents. This anticipation, in turn, is more effortless when agents share beliefs. When agents share beliefs, they anticipate counter-arguments more easily and, as a consequence, they do not publish arguments that their communities would immediately rebut. Indeed, by doing so an agent avoids the loss of reputation that would result of not being able to defend her own stance.

Therefore, shared beliefs may act as shared standards of evaluation, that is as a mechanism to weed out bad arguments from good arguments. By way of illustration, we can compare the social sciences with the natural sciences. The latter is characterized by a higher degree of shared beliefs than the former, and, consequently, agents are better at anticipating what the possible counter-arguments to their own arguments will be. This implies that the standards for arguments are clearer and sharper, and less weak arguments are published. In the context of science such standards are tracked by communal practices such as peer review.⁷

In our model, we investigate the role of shared beliefs insofar as they act as shared standards, that is as a filter that enables a community to reject arguments which are too weak. In this way weak arguments do not affect scientific discourse. That is, before a newly generated argument is added to individual views, it is decided whether it is worthy of consideration. If not, only the author adds the argument to her own view, while no other agent has the opportunity to integrate it into their view.

The probability for an argument a to be discarded depends on its strength $S(a)$ and a fixed shared belief parameter $shBel \in [0, 1]$:

$$p_{discard} = shBel \cdot (1 - S(a)).$$

The higher $shBel$, the higher the probability that weak arguments are filtered; for $shBel = 0$, we are in a community without shared beliefs.

⁷For instance, the variability of the quality of grant proposals (as judged by the reviewers) submitted to the Austrian Science Fund (between 1999 and 2009) is twice as high in the social sciences than in the natural sciences (Mutz et al., 2012). Similarly, the reviewers' ratings (in terms of their inter-rater reliability) vary considerably more in the former when compared to the latter field.

In this way our model implements a "slim version" of the mechanism of shared beliefs proposed by Mercier & Heintz. One that focuses on the collective effect of shared beliefs, while bracketing the individual processes (such as anticipations and reputation) of the agents. By doing so, we make sure to introduce in the model the kind of effect on argument production Mercier and Sperber had in mind, that is that of an epistemic collective filter.

7.4 Simulations and Observations

In our simulation, a community of agents investigates two research programs by producing new arguments. Each research program has an objective strength, which corresponds to the strength of each of its central arguments. Although the agents do not know the strength of either of the programs, they aim at understanding which program is stronger by debating and arguing about the central arguments. In fact, the strength values of the central arguments affect the shape of the debate, insofar as stronger arguments are harder to attack, and usually easier to defend. By generating arguments, agents aim at eliciting which program is objectively stronger. In an epistemically ideal situation an objectively stronger research program would always be the one with the highest score. Yet, this may not always be the case, e.g. when agents are biased.

Each scientist aims to support the stronger research program: for this, they try to attack different arguments and choose to support the research program with the higher score (to them). We hypothesize that in baseline conditions, the scores the scientists assign to the two programs will likely respect the respective strengths of the central arguments.

Now that the protocol and the way opposing views have been presented we shall summarize the parameters of our simulation (see Table 7.1 for an overview). To do so, we formally define the community *Com* and environment *Env*. A community *Com* describes all the socio-epistemic features of the collective who engage in the debate: its size (N), the initial support granted to *ResProg*₁ (n_1^0), the strength of the bias (*bias*), and the strength of the shared beliefs (*shBel*). It is a quadruplet $Com = \langle N, bias, shBel, n_1^0 \rangle$. An environment *Env* describes the conditions of the scientific controversy: the number of central arguments per research program and the coefficient that determines the standard deviation of the distribution from which the strength of a new argument is drawn. It is given by a pair $Env = \langle N_{CA}, StDev \rangle$. Finally, there are the two research programs, *ResProg*₁ and *ResProg*₂, that come with their respective strengths S_1, S_2 .

Table 7.1: Parameters for the simulation of our model.

Parameter	Definition	Test Interval	Default Value
N	Number of agents.	$[2, 12]$	10
n_1^0	Initial support for $ResProg_1$.	$[0, N]$	5
N_{CA}	Number of central arguments in each research program.	$[1, 10]$	5
p_{see}	Probability for an agent to add an argument to her personal view.	$[0.5, 1]$	0.5
$StDev$	Standard deviation of the probability distribution for the generation of argument strength.	$[0.1, 0.5]$	0.2
$bias$	Bias parameter.	$[0, 1]$	0.3
$shBel$	Shared beliefs parameter.	$[0, 1]$	-
S_1	Strength of $ResProg_1$.	$[0, 1]$	-
S_2	Strength of $ResProg_2$.	$[0, 1]$	-

7.4.1 Epistemic evaluation

We want to characterize the **epistemic success** of a community Com in a specific environment Env . The idea is to measure how good the collective of agents is at selecting the best research program for different values of the strengths of the research programs (S_1, S_2) .

To evaluate the performance of a community we start by defining two fundamental notions. First, we define the agentive support, in short **support**, of a research program $ResProg_i$ at time t , i.e. n_i^t , as the number of agents preferring $ResProg_i$ at step t . As n_2^t can be easily inferred from n_1^t (if an agent does not support one research program, she supports the other), we will only refer to n_1^t from now on. Intuitively, the greater the support of the strongest research program, the greater the epistemic success of the community. From our first analysis of the simulations, the most relevant factor is the *difference* between each research program's strengths. To characterize this intuition, we define the **strength difference** $d \in [-1, 1]$ as the difference between the strengths of the two research programs ($d = S_1 - S_2$ with S_1 and S_2 the respective strengths of $ResProg_1$ and $ResProg_2$). Such a difference is positive when $ResProg_1$ is stronger than $ResProg_2$, and conversely.

A very clear indicator of the epistemic success of a community would be if every member always preferred the strongest research program. Yet, since our model is stochastic, an otherwise rather accurate community may sometimes end up supporting the weaker program. To have a more relaxed criterion of success, we define an epistemically successful community as a community where, on *average*, a majority of agents prefers the strongest program. An epistemically successful community is

one where the average support for the stronger program is higher than half of the community.

We report the *average* support for a research program over multiple runs with the same strength difference d . We launch 10 000 simulations with the same parameters for Com and Env ; the strength of the research programs is randomly generated. Then, we define 40 distinct values of d covering the $[-1, 1]$ interval, each distant from 0.05, and round the strength difference of each simulation to the closest of these values: we end up with roughly 250 runs for each value of d . For each of these runs we measure the support of each research program at 100 steps (i.e. at $t = 100$), which is a reasonable convergence time, as explained in Section 7.5.1. Accordingly, we denote $\overline{Supp}_d$ as the average support of a community Com at step 100 for a specific value of d . We indicate with SE the standard error over the value of $\overline{Supp}_d$. The standard error gives us an estimation of how $\overline{Supp}_d$ could vary if we were to repeat the experiments. Notably, by averaging the value of support over 250 runs per parameter combination, we make sure the standard error is consistently low.

Definition 29 (Epistemic Success). *We say that a community Com is **epistemically successful** for a strength difference $d \in [-1, 1]$ iff the following two conditions hold:*

1. $\overline{Supp}_d + SE > N/2$, if $d > 0$; and
2. $\overline{Supp}_d - SE < N/2$, if $d < 0$.

That is, a community Com is epistemically successful for a strength difference $d \in [-1, 1]$ if and only if the average support it produces for the stronger research program is higher than half the size of the community. We say that Com is epistemically successful for an interval $[d_1, d_2]$ and for an environment Env iff it is epistemically successful for all $d \in [d_1, d_2]$ and for Env .

We introduce intervals in order to compare the performance of different communities. Intuitively, a community Com_1 is more epistemically successful than another community Com_2 whenever the interval in which Com_1 is epistemically successful is larger than the interval in which Com_2 is so. Indeed, some communities grant a higher support to the strongest program only when the strength difference is large, while do not do so in other cases.

As a consequence, we say that a community is **fully epistemically successful** if for any interval $[d_1, d_2]$ and any environment obtained by the combination of values in Table 7.1 such community is epistemically successful.

Finally, we define the **polarization ratio** of a community as the ratio between the largest group of agents and the smallest group, i.e.

$$PolRat^t = \frac{\min(n_1^t, n_2^t)}{\max(n_1^t, n_2^t)}. \quad (7.3)$$

Following [Bramson et al. \(2017b\)](#)'s classification, this is what is called a *size parity* based form of polarization. The polarization ratio will help us delineate a specific bias feature in Section 7.5.3. For the same reasons highlighted above, we compute the average polarization ratio $\overline{PolRat}$ by averaging over the values for polarization ratio at 100 steps of 250 runs per parameter combination.

7.5 Results

The main results we observe are the following.

1. Non-biased communities are good at selecting the best alternative (i.e. fully epistemically successful).
2. Biased communities
 - are less epistemically successful than non-biased ones, but
 - tend also to polarize less than non-biased ones.
3. The negative impact of the bias on epistemic success can be mitigated by:
 - An equal number of agents supporting each alternative in the beginning.
 - A mechanism to filter out weak arguments at the community level, the "shared beliefs".

With the former being more efficient than the latter.

Unless specified otherwise, all the following plots and observations are made using the default values for the parameters in Table 7.1.

We divide the present section as follows. First, we discuss the asymptotic behaviour of our simulation, and we argue that measuring the average support at step 100 is a reasonable choice. Then, we briefly study the behaviour of a non-biased community, and we compare it with that of a biased one. Finally, we show the way shared beliefs may mitigate the negative impact of the bias.

7.5.1 Time convergence

In this section, we study the asymptotic behaviour of communities. Although some communities may reach a stable situation in terms of support for each research program, this is not the case for every parameter combination. For certain sets of parameters, no situation is stable, since there is always a non-zero probability that

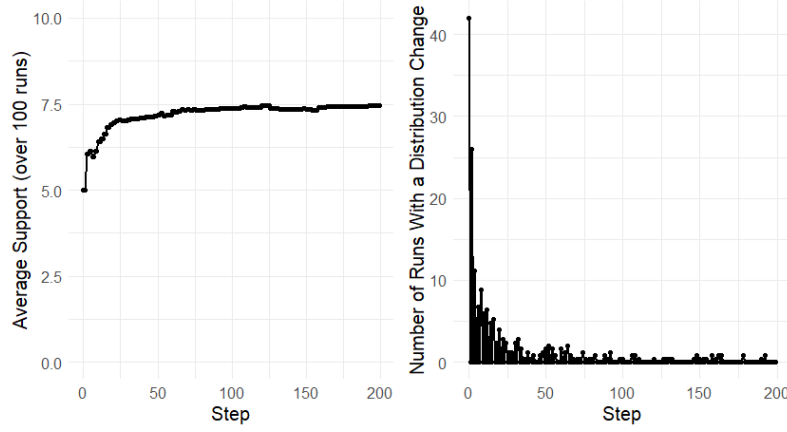


Figure 7.6: On the left: the evolution of the value for the average support for $ResProg_1$ is plotted at any step. On the right: the number of runs (y-axis) in which the value of the support changes in the time-step represented on the x-axis. The total number of runs is one hundred in both cases. Here, $d = 0.1, n_1^0 = 5$. Similar plots can be obtained for the whole parameter space.

agents generate a set of arguments that lead one or more agents to change their preference(s).

Regardless, the number of runs in which at least one agent changes her mind at a given time-step decreases with every step (see Figure 7.6). And, notably, after 100 steps, almost no run undergoes any change. For this reason, we choose to evaluate the performance of a community at that time. ⁸

In general, as time goes on, agents become less and less likely to change their minds for two reasons: first, as the agents' views become more and more different, every agent adds to her view fewer arguments per step, as it is very likely new arguments attack arguments she is not aware of. Second, as the agents' views become larger and larger, the addition of a new argument becomes less and less likely to change their minds.

7.5.2 The baseline model

We now turn to analyze the baseline model, i.e. a non-biased community with no shared beliefs ($bias = 0$ and $shBel = 0$). By testing its results against our intuitions, we can ensure our modeling choices reflect our conceptual understanding of the matter. In particular, we expect such a community to be fully epistemically successful: that is, we expect most agents to prefer the stronger research program. Similarly, we expect the average support for $ResProg_1$ (that is $\overline{Supp_d}$) to increase as the difference in strength between the two research programs ($d = S_1 - S_2$) increases.

⁸We checked that this is the case in communities as large as 100 agents, and did not find any effect of the size of the population on the results.

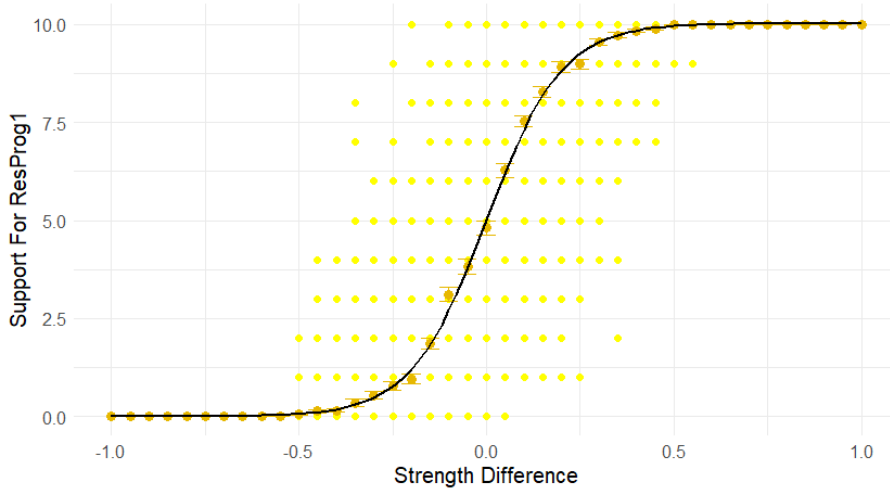


Figure 7.7: The support for $ResProg_1$ is plotted against the strength difference d . The light small yellow points represent the value of support $Supp$ for singular runs. The other ones are the values for the average support ($\overline{Supp}_d$). The standard error is represented as an interval around $\overline{Supp}_d$. The black line is the graph of the function $y = 10(1 + e^{-9.3d})$, which is the best approximation of our distribution.

Unbiased communities perform in accordance with our intuitions, as they are fully epistemically successful, and their average support for $ResProg_1$ ($\overline{Supp}_d$) increases as d increases. As can be seen in Figure 7.7, the average support increases as the strength difference increases and is higher for $ResProg_1$ when the difference is higher than zero ($d > 0$): when $ResProg_1$ is stronger, a majority of agents prefer it. In addition, when the difference in strength is large in absolute value ($|d| > 0.5$), the community always reaches a correct consensus: baseline communities are strictly epistemically successful on the interval $[-1, -0.5] \cup [0.5, 1]$. Important to note is that the value of S_1 and S_2 do not matter: only the difference of strength between programs determines the number of agents on each side at the end.

The individual data points may be found quite far away from the averaged values (Figure 7.7). when the absolute value of the difference is small ($|d| < 0.5$), $Supp$ takes on many different values depending on the run. Instead, when the absolute value of the difference is large ($|d| > 0.5$) the support is either zero or ten, meaning the community always reaches a consensus (remember that the size of the community is ten). In fact, the greater the difference, the more the agents are pushed towards the strongest theory, up to the point in which a consensus on it is the only distribution likely to last.

The relationship between the average support ($\overline{Supp}_d$) and the strength difference (d) shown in Figure 7.7 can be approximated by a *sigmoid* model with formula:

$$\overline{Supp}_d = \frac{\alpha}{1 + e^{-\beta \cdot d}} \quad (7.4)$$

with $\alpha = 10, \beta \in [5, 11]$.⁹ The value α is the maximum of the curve: it is equal to

⁹For the default values, $\beta = 9.3$ and the residual standard error is 0.1045 on 39 degrees of

the total number of agents, whereas β controls the curve's steepness and depends on the value of the parameters (Table 7.1). We consider the emergence of a sigmoid distribution significant, as it underlines crucial features of our model.

1. The community is highly sensitive to any change in the strength difference when such difference is small ($|d| < 0.25$), that is to small changes in the strength difference correspond large changes in the average support.
2. The community is less sensitive to any change in the strength difference when such difference is rather large ($|d| > 0.5$). The community always converges to consensus whenever the strength difference is rather large, as all agents choose the strongest theory.

The emergence of the sigmoid is robust for any parameter combination in Table 7.1, i.e. it is always possible to fit the sigmoid function effectively to describe the behaviour of the average support ($\overline{Supp_d}$). However, while α is stable and equal to the number of agents ($\alpha = N$) the value for the steepness of the function (controlled by β) changes with respect to how the parameters change (see Appendix B).

To provide the reader with an example, we study how the steepness of the sigmoid (β) is influenced by the probability for agents to add an argument to their own views (p_{see}). As such probability decreases, the central part of the sigmoid becomes less steep. This implies that the community is less sensitive to small changes in the strength difference in such a part. Decreasing the probability of adding an argument causes the agents to be less effective in identifying the stronger theory. Indeed, the more arguments are included the more their distribution in the agents' views is likely to reflect the strength difference between the two research programs. If an agent only includes a part of all the produced arguments in its view, that part is more likely to be misleading.

In summary, our baseline model fits our conceptual expectations, which satisfies a basic requirement for its representational adequacy. In addition, it also features the emergence of a specific pattern, i.e. a sigmoid relationship between the average support and the strength difference, which highlights some interesting details of the performance of the community.

7.5.3 A biased community

Now that the behaviour of a non-biased community is clear, we proceed to answer our first research question, concerning the effect of myside bias (Section 1), by discussing the three following findings.

freedom, i.e. 39 different values for d : 0, 0.05, 0.1, 0.15, ...1.

1. The bias has a detrimental effect on communities, as biased communities are often not fully epistemically successful. In particular, biased communities tend to preserve the *status quo* more than unbiased communities.
2. Equally distributed biased communities are fully epistemically successful, i.e. they neutralize the detrimental effect of the bias.
3. Biased communities tend to be less polarized than unbiased ones.

7.5.3.1 The detrimental effect of the bias

Biased communities are usually less epistemically successful than non-biased ones. A biased community mainly differs from a non-biased one because the initial support for the research programs, i.e. how the agents are divided initially in terms of preferences, influences the evolution of agents' preferences. In a non-biased community the initial support does not matter. In a biased community, if a majority of agents support *ResProg*₁ initially ($n_1^0 > \frac{N}{2}$) the average support for *ResProg*₁ will be higher than if the community were not biased, while if the initial support for *ResProg*₁ is a minority ($n_1^0 < \frac{N}{2}$), the final average support for *ResProg*₁ will be lower (see Figure 7.8). Therefore, if there is a clear majority preferring one side at the beginning of the debate ($n_1^0 \neq \frac{N}{2}$), a biased community is not fully epistemically successful: the average support will be higher for the weaker research program for a certain interval of values of the strength difference.¹⁰ For example, if nine out of ten agents start out preferring *ResProg*₁ (as in the green line in Figure 7.8) the average support is higher for *ResProg*₁ even if *ResProg*₂ is stronger than *ResProg*₁ of a value 0.2 (that is when $d = \text{ResProg}_1 - \text{ResProg}_2 = -0.2$).

In particular, given a certain value for the bias, the further away is the initial support for *ResProg*₁ from half the size of the community, the smaller the interval in which a community is epistemically successful, as shown in Figure 7.8. In this sense, we say that biased communities tend to preserve the *status quo*, as they tend to provide more support for the program that was preferred in the beginning. A similar mechanism is observed for the value of the bias: the greater the bias, the smaller the epistemically successful interval, as the more the average values depart from the baseline results (see the comparison with $shBel = 0$ between $bias = 0.3$ and $bias = 0.5$ in Figure 7.10).¹¹

¹⁰We observed this behaviour for a bias as lower as 0.05. When the bias gets lower than that value, it is hard to distinguish whether or not there is any statistical difference between a situation with $bias = 0$.

¹¹Although this behaviour can be observed through the whole parameter space, i.e. for any possible parameter combination studied (see Table 7.1), it is worth highlighting the slightly different behaviour of what we call *Limit cases*. As the probability of attacking a research program cannot exceed 1, sometimes the bias may provide a 'bonus', that is smaller than its numerical value. Consequently, it may be the case that one research program is more 'helped' by the bias than the other. Although such detail does not have a major impact, the results change slightly when the bias is strong. In particular, it can be the case that when $n_1^0 = 5$, the average values do not

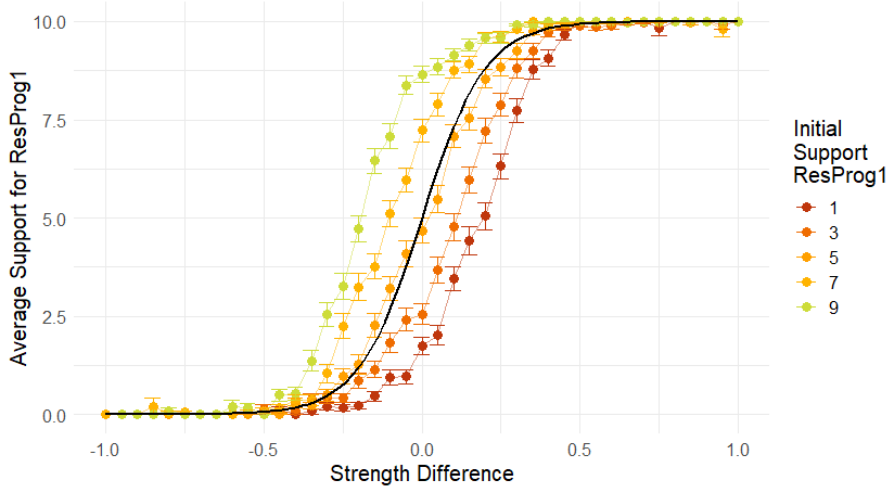


Figure 7.8: The average support for $ResProg_1$ is plotted against the difference d . The agents all have a bias of 0.3. The colours represent initial support for $ResProg_1$. The standard error is represented as an interval around $\overline{Supp}_d$. The black line represents the average support produced by the baseline model (Equation (1)).

The results of a biased community can also be approximated by Equation (1). To do so, we introduce a generalized notion of strength difference, which allows us to use Equation (1) to fit the results of biased and unbiased communities. We call this refined notion of strength difference *biased strength difference*.

Definition 30 (Biased Strength Difference). *In the context of a debate with parameters S_1, S_2, n_1^0 , the biased strength difference is defined as : $d_{bias} = n_1^0 N \cdot (\min(S_1 + bias, 1) - \max(S_2 - bias, 0)) + (N - n_1^0) N \cdot (\max(S_1 - bias, 0) - \min(S_2 + bias, 1))$.*

The biased strength difference is derived from the initial distribution of agents, their bias and the strength of the two research programs. It can be understood as a weighted average of both groups' probabilities of generating attacks against each research program at the beginning of the debate. The biased strength difference generalizes the strength difference since, in case $bias = 0$, we have $d_{bias} = d$.

Observation 1. *Using the biased strength difference (d_{bias}), we can describe all the results obtained so far (that is, those in Figure 7.8 and Figure 7.7) as follows:*

$$\overline{Supp}_{d_{bias}} = \frac{\alpha}{1 + e^{-\beta \cdot d_{bias}}}, \quad (7.5)$$

where, again, $\alpha = 10$ and β depends on the parameters of the debate setting.¹²

This is remarkable for two reasons. On the one hand, it is interesting that despite the combinatorial complexity of the protocol, we can derive the average

precisely follow the baseline model. For example, assume $d = 0.25$ and $bias = 0.4$; if $S_1 = 0.8$ and $S_2 = 0.55$, the average support for $ResProg_1$ is lower than if $S_1 = 0.6$ and $S_2 = 0.35$. From now on, we call a situation where the sum exceeds 1, a *limit case*.

¹²Again, for the default values, $\beta = 9.3$ and the residual standard error is 0.76 on 39 degrees of freedom, i.e. 39 different values for $d_{bias} : 0, 0.05, 0.1, 0.15, \dots, 1$.

support using a simple formula, which only depends on the initial conditions. On the other hand, it allows us to understand the behaviour of a biased community with the same mechanism as an unbiased one. A biased community dealing with two research programs with strength S_1 , S_2 and with a certain biased strength difference (d_{bias}) produces the same average support as a non-biased community in the same environment dealing with two research programs with strengths S'_1 and S'_2 such that its correspondent strength difference is the same of the biased strength difference ($d' = d_{bias}$). It is as if a biased community perceives the two research programs as having different strengths, which are influenced not only by the bias but also by the number of people supporting each theory initially.

7.5.3.2 Equally distributed biased communities

Our results show that the community *Com* is fully epistemically successful when there is an equal number of agents preferring each research program at the beginning ($n_1^0 = n_2$). We call these communities "equally distributed", referring to their *initial* distribution of support. Consider again Figure 7.8: when the initial support for *ResProg₁* is five ($n_1^0 = 5$), the final average support is the same as when the community is not biased.¹³ This suggests that an equal initial distribution of agents cancels out the impact of the bias.

These results align with the conjecture by [Mercier and Heintz \(2014\)](#) that collective reasoning is less affected by the detrimental impact of myside bias than the individual one. In our model, if a community is composed only of one individual, it would never be epistemically successful (as there is no way to distribute equally only one agent); on the other hand, a biased community of more than one scientist may be equally distributed and thus, epistemically successful. Such an observation further validates our simulation.

In summary, biased communities are not epistemically successful, as the bias makes communities more inclined to produce a higher average support for the research program that was more supported initially, even if it is not the objectively stronger one. Such an effect is positively correlated with the intensity of the bias; yet, quite surprisingly, it does not come into play if the community is initially equally distributed.¹⁴

¹³This is the case under the assumption that we do not incur in 'limit cases' - see Footnote 11)

¹⁴[Appendix C](#) analytically investigates the behaviour of a community of two agents. This may help understanding what makes equally distributed biased communities behave exactly like non-biased ones.

7.5.3.3 More bias, less polarization

Finally, we also discuss the effect of bias on the polarization of a community. As can be noted from Figure 7.9 more biased communities tend to polarize less, that is to exhibit a lower average polarization ratio. Although this result may perhaps seem

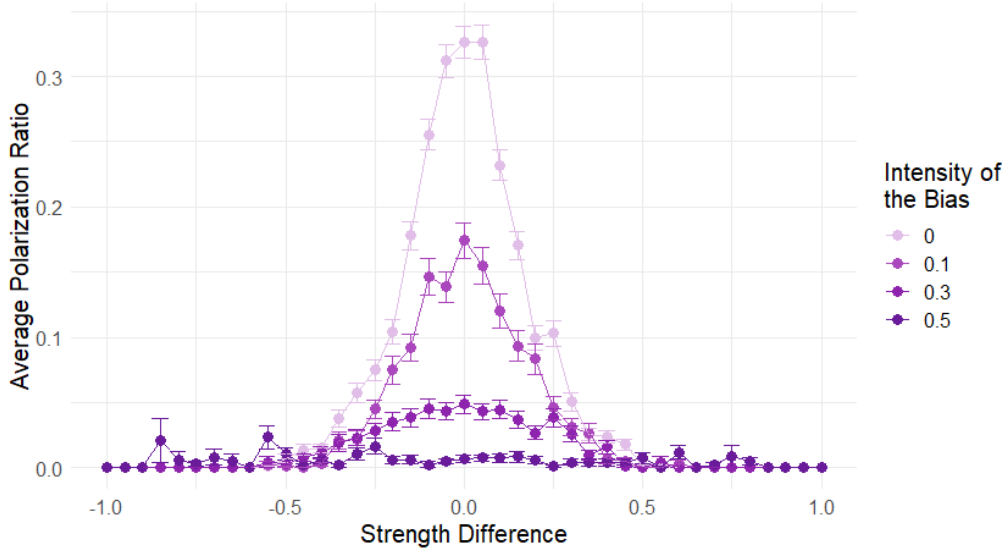


Figure 7.9: The average polarization ratio plotted against the strength difference d . The colours represent different values for the bias.

surprising, it is easily explained. Biased communities are not more likely than non biased ones to form a consensus. Yet, when the agents of a biased community end up preferring all the same program they are very likely to not change their preferences ever again. Indeed, in such a situation agents are highly unlikely to produce enough arguments against the community current preferred research program to change someone's mind. As a consequence, even if non-biased communities are equally likely to form consensus as biased ones, consensus is way more stable in the latter. This explains why biased communities exhibit a lower average polarization ratio.

7.5.4 Results with shared beliefs

To check the effectiveness of the shared beliefs filter, we study how a biased community with a certain degree of shared beliefs compares to a baseline community: the closer the behavior of the biased community to that of the baseline community, the more the shared belief filter has a mitigating effect.

We find that shared beliefs do reduce the detrimental impact of the bias, but this effect is quite limited. To see this consider the lower line of plots of Figure 7.10, where we represent the average support of different rather biased communities ($bias = 0.5$) with respect to the average support of a similar non-biased community.

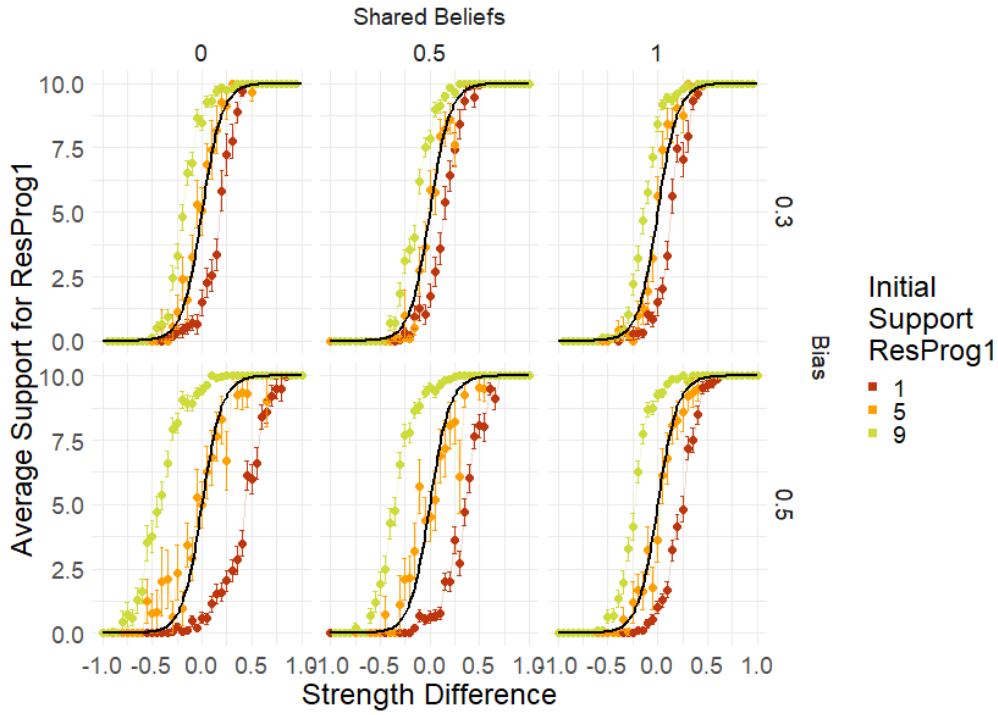


Figure 7.10: The average support for $ResProg_1$ plotted against d . The colours represent the initial support for $ResProg_1$. The standard error is represented as an interval around the $\overline{Supp}$. The subplots corresponds to different intensity of shared beliefs and of the bias. The black line represents the average support produced by the baseline model (Equation (1)).

The red points describe the average support of a biased community with an initial support of one ($n_1^0 = 1$), the green ones the average support of a biased community with initial support of nine ($n_1^0 = 9$), and finally the black line represents the sigmoid approximation of the average support in a similar non-biased community. As the shared beliefs parameter $shBel$ increases, the average supports of the biased communities get closer to this baseline approximation: a rather biased community ($bias = 0.5$) is epistemically successful in a greater interval for d when $shBel$ increases. Yet, even when the degree of shared beliefs is at its maximum, the initial support of $ResProg_1$ still influences the final average support ($\overline{Supp}$) as the community is not yet completely epistemically successful (which can be seen by the fact that the green and red points do not follow the black sigmoid).

In addition, when the bias is quite low, the effect of the shared beliefs is almost invisible: consider the first line of Figure 7.10, where there is none to little difference between the three plots. Since the shared beliefs lead to a suppression of weak arguments, their mitigating effect on the bias works best when the bias allows agents to produce a lot of weak arguments. This is not the case when the bias is small.

Similarly, shared beliefs are ineffective in mitigating the bias's impact when the difference between the strength of the two research programs is very small (d close to 0). This can be seen in Figure 7.10, as the distance between the average support for the biased community and the sigmoid approximation does not diminish much

in the central area of each plot, that is when the difference is the smallest. Shared beliefs help agents to discriminate between good and bad arguments, but in such area all arguments have more or less the same strength. Consequently, in those situations, shared beliefs are not very effective.

7.6 Related Works

We now provide a brief comparison of our approach and extant studies while underlining the novelty of our model. Although no previous ABM has studied myside bias as a production bias in the context of argumentation, many ABMs have been proposed to either study cognitive biases in the context of scientific inquiry or opinion dynamics in argumentative contexts. We compare our work to both these two streams of literature.

As we presented in Chapters 3 and 4, previous ABMs based on argumentation have been used to study opinion dynamics and polarization effects in argumentative settings (Gabbriellini and Torroni, 2014; Butler et al., 2019; Taillandier et al., 2021; Mäs and Flache, 2013; Proietti and Chiarella, 2023; Singer et al., 2018), gradual opinion in debates (Dupuis de Tarlé et al., 2022a), network effects in the context of scientific inquiry (Borg et al., 2018, 2017b), or argumentation-based procedures for how to choose among scientific theories (Borg et al., 2019). In terms of modeling framework, the main differences between our model and this stream of literature are mainly two.

First, our model allows for a more natural process of argument generation: any agent can investigate any argument, and possibly attack it. Attacks may be weak or strong depending on where they are directed to. This represents the features of a deliberative process, in which the discussants can create arguments that may not always be *convincingly strong*, and also allows us to include myside bias as a production bias easily. This is not the case for either of the past models, in which either the arguments are a finite set and can be learnt just by meeting someone who already possesses them (Mäs and Flache, 2013; Dupuis de Tarlé et al., 2022a; Proietti and Chiarella, 2023; Singer et al., 2018), or the arguments are a finite number and can only be *discovered* as pieces of evidence by scientists who look into the right place (Borg et al., 2018, 2019).

Second, our model introduces a new mechanism to evaluate the epistemic outcome of the discussion. This is, again, a novel feature that more traditional models do not usually have, as they are not concerned with the epistemic impact of a mechanism but rather with the dynamics it generates.

Although no model allows for the representation of myside bias as a production bias, computational models have been used to capture the effects of cognitive biases intended as *evaluation biases*. (Liu et al., 2015; Fu and Zhang, 2016; Banisch and

Shamon, 2021; Proietti and Chiarella, 2023; Dupuis de Tarlé et al., 2022a) all model biased agents as agents who are more likely to ignore arguments against their present position than arguments in favour of it. Such agents exhibit an evaluation bias, insofar as they fail to evaluate arguments impartially. In addition, (Gabriel and O'Connor, 2022) and (Baccini and Hartmann, 2022; Baccini et al., 2023) model evaluative biases in a Bayesian framework. They consider the exchange of arguments as an exchange of evidence, and understand biased agents as agents who discount evidence against their own current position. Finally, in (Proietti and Chiarella, 2023) and (Singer et al., 2018) biased agents tend to forget arguments opposed to their present views sooner than the other ones.

Most of these models find that biased communities polarize more often or that polarization is usually expected to last longer. In contrast to them, our results suggest that if myside bias is understood only as a production bias, i.e. a tendency to produce arguments only in favour of one's view, biased communities are less likely to be found in a polarized state. This suggests that polarization is affected in different ways by various forms of myside bias. However, a more thorough comparison of the results of these two types of myside bias is needed before drawing any conclusion, as our results only take into account one of the possible measures of polarization (Bramson et al., 2017b), and a specific setting in which the bias can emerge.

Among the many articles that study the effect of biases in deliberative settings, (Gabriel and O'Connor, 2022; Baccini et al., 2023) are the only ones who look at the epistemic performance of a community. Baccini et al. (2023) provide a bayesian interpretation of myside bias as an evaluation bias, based on their previous work (Baccini and Hartmann, 2022), and explore under which conditions biased agents outperform non-biased ones. They find that this is the case mostly when the size of the deliberating community is small and agents start the discussion with already a favourable prior belief in the correct hypothesis. Gabriel and O'Connor investigate the effect of bias for scientific inquiry whenever scientists need to choose which theory to use among two. They show that moderate levels of confirmation bias, as a skewed evidence evaluation, can help scientists to reach an accurate consensus more often.

7.7 Conclusion and Outlook

We used multi-agent modeling and abstract argumentation to represent a community of scientists debating two distinct research programs. We tested our model against general conceptual expectations and registered the emergence of a specific relationship between the average support for a research program and the strengths of the two research programs. Namely, the greater the difference of strength between the two research programs, the more agents will successfully prefer the stronger one.

Following Mercier and Heintz (2014) and Mercier (2016, 2017), we introduced

the possibility for agents to be biased when producing new arguments. We measured the detrimental effect of myside bias by showing that biased communities are usually not fully epistemically successful, as they may produce higher average support for a weaker research program if such a research program had more initial support. We discovered that this harmful effect of the bias is not present in communities that are equally distributed, that is when a similar number of agents supports each research program at the beginning. In addition, we also showed that biased communities tend to polarize less than non-biased ones. This result stands in sharp contrast with results concerning more classical types of ‘evaluation biases’, such as those obtained by [Mäs and Flache \(2013\)](#); [Proietti and Chiarella \(2023\)](#); [Singer et al. \(2018\)](#) who find that biased communities polarize more.

Finally, we have demonstrated that communities with more shared beliefs for the evaluation of arguments can mitigate the detrimental effect of the bias. However, the beneficial effect of shared beliefs cannot completely cancel the impact of bias, and becomes rather negligible when the two research programs have a similar strength. In this way we confirmed and qualified a conjecture by [Mercier and Heintz \(2014\)](#).

In light of these results, our contribution is twofold. On the one hand, we have provided a simulation-based test for previously proposed hypotheses, such as those by [Mercier and Heintz \(2014\)](#). On the other hand, we have made a first step in offering a more systematic understanding of the impact of myside bias on the collective performance of scientists.

Our findings hinge on a number of assumptions. First, we assume the objective strength of the research programs to not be visible to the agents, but to only influence the process of argument generation. Accordingly, we measure the performance of a scientific community based on its ability to identify the strongest between two scientific programs. We assume the two scientific programs to have the same number of central arguments: each program has a specific strength and this strength is assigned to each of its central arguments. Agents aim at eliciting which program is the strongest by looking for counter-arguments to central arguments, or by attempting to defend them. This idealized picture of scientific debate allows us to focus on the effect of myside bias, and bracket many of the complex details of scientific inquiry. However, future work may be directed at relaxing these assumptions and study the robustness of our results with a more nuanced representation of science. For example, a research program with a significantly higher number of central arguments than the other may make the task of scientists easier – and, consequently, mitigate the detrimental effect of the bias.

We also represent scientists as impartial in evaluation arguments and high performing agents. They remember every argument they learn, and, they always support the most promising and solid theory, regardless of whether they are biased or not. This assumption captures Mercier and Sperber’s account of myside bias, which they take to not affect argument evaluation: agents are always willing to change their minds when presented with enough arguments ([Mercier and Sperber, 2017](#)). If

agents were more steadfast in their preferences, the bias may cause a more polarized situation, and the initial distribution of agents may have an even stronger impact.

Although our model aims at representing scientific debates, it can be easily applied to any debate in which these assumptions are reasonable. Consider, for example, a committee of urban planners discussing two possible ways of building a stadium. Under the assumption that there is an objectively superior option, e.g. because it would make spectators safe, or would make the energetic system more efficient, one could model the discussion leading to their decision in our framework.

Another way to interpret the results of our models is through the lens of argumentative strategies. Biased communities defending weaker research programs tend to produce numerous weak arguments. Shared standards help mitigate this bias by acting as a filter, reducing the number of weak arguments that reach the public sphere. This scenario is reminiscent of the 'Gish Gallop' strategy, named by [Scott \(1994\)](#) after a well-known climate skeptic. The Gish Gallop involves overwhelming opponents with a stream of arguments, regardless of their strength.

Our study opens ample space for future work. While we have shown that equally distributed groups may be able to cancel the detrimental effect of the bias, new questions arise. On the one hand, which incentive structures can help communities obtain such a division of labor? On the other hand, are there other mechanisms that lower the detrimental effect of myside bias? [Mercier and Heintz \(2014\)](#) propose a high degree of interactivity of scientific communication to overcome fragmentation of argumentative exchanges (p. 521). A possible approach is to link this feature with the work on adversarial collaboration by [Clark and Tetlock \(2021\)](#).

Another open question concerns the effectiveness of shared standards of argument evaluation in our model, which are helpful only if one research program is clearly stronger than the other. This raises the question: can shared standards also help communities if the rival research programs are similar in strength, and if so, how? Are there some additional mechanisms that support the mitigating role of shared standards?

As noted by [Mercier and Heintz \(2014\)](#) and [Dutilh Novaes \(2018\)](#), when a myside biased agent fails to predict potential defeaters to one of her produced arguments, she risks not being able to reply to criticism. This in turn may negatively affect her reputation. Reputation management and decision making under uncertainty are thus important features to consider in future iterations of our model.

CHAPTER 8

A Brief Synthesis of the Models

Both Chapter 6 and Chapter 7 present agent-based models of scientific communities where abstract argumentation is used to represent agent cognition. Here, I present a brief summary of the main differences between the two agent-based models and compare their results.

Table 8.1 presents a side by side comparison of the main features of the two models. Table 8.2 presents a summary of the results obtained by each model. Due to differences in the frameworks, protocols, and research questions of the models, some results are difficult to compare directly. Specifically, the findings about biased communities in Chapter 7 are challenging to relate to those in Chapter 6, where no similar bias mechanism is implemented. However, it is noteworthy that in both models, agents within non-biased communities achieve a satisfactory level of epistemic success. Additionally, when observing the outcome of the model of Chapter 7, we observed qualitatively that communities of agents tend to stabilize more quickly when the difference in strength between the two research programs is significant, while greater variation in agents' preferences occurs when the difference is minimal. This pattern resembles the findings in Chapter 6, where agents' opinions stabilize slowly when the issue has a truth value close to 0.5. In both cases, it seems that the agents have a harder time forming stable positions when the debate centers around questions that can be described as "undecided".

Epistemic success is examined in both models. Based on the findings, one might conclude that the ideal scenario for maximizing epistemic success is a community of non-biased, similar, and highly connected agents. However, caution is needed when interpreting the impact of diversity on epistemic success, as we showed that this outcome is tied to the specific argument generation protocol used in Chapter 6. This nuance explains the apparent contradiction with the results from Chapter 7, which indicate that having maximum diversity at the start of the debate—meaning an equal number of agents favoring each research program—enhances the epistemic success of biased communities.

A promising direction for future research would be to introduce a form of myside bias into the model from Chapter 6. This would allow us to verify the findings from Chapter 7 and, if discrepancies arise, identify which aspects of the models are responsible. As discussed in Chapter 3, comparing different implementations of the same mechanism is a valuable epistemic practice that can either reinforce or challenge the results of a model. This approach would deepen our understanding of the dynamics induced by myside bias within scientific communities.

The research question that led me to model communities of scientists was that

of the study of socio-technical controversies. These two models contribute to answering part of that question, as they allow us to better understand the dynamics at work in debates between scientists, and, more generally, agents looking for the truth. In particular, they illustrate the importance of social structures on the outcome of deliberation, whether it is the communication networks within which agents evolve, the relative number of agents on each side of the debate or the presence of community-based filters. The model presented in Chapter 7 proposes an illustration of what happens when biased agents defend a weak theory (if we model them according to Mercier and Sperber's hypothesis): they produce many arguments of low value.

Building on these models, one could create a comprehensive ABM representing all the actors in a socio-technical controversy. This would enable an exploration of how social structures influence the circulation of arguments and shape the outcomes of these controversies. Specifically, it could provide insights into how media coverage filters scientific arguments and sways public opinion. We could also represent other forms of social structures within scientific communities, such as hierarchies between agents or a reputation mechanism.

	Chapter 6	Chapter 7
Arg. graphs	Issue-oriented bipolar graphs.	Attack-only graphs with a specific structure: a collection of independant sub-debate which are issue-oriented with "central arguments" as the roots.
Arguments	Arguments are characterized by truth values.	Arguments are characterized by strengths.
Semantics	The LSS gradual semantics (see Chapter 5).	The grounded semantics (but any set-based semantics is equivalent).
Agent's position	The agent's opinion is the acceptability score of the issue.	The agent's preference is the research program with the highest number of central arguments which are acceptable.
Argument generation	Depends on an argument's truth value and its score in the agent's mental graph.	Depends on the argument's strength and the bias of the agent.
Learning arguments	Adding an argument and all its relations to the agent's mental graph.	Adding an argument and all its relations to the agent's mental graph.
Social features	The agents are linked by a social network.	There is a "shared standard" which acts as a filter, preventing certain arguments of being made public.
Sharing arguments	The new argument is shared with the agent's direct neighbors in the network.	If they pass the filter, the new argument is shown to all agents. Each agent has a certain probability to learn the new argument.
Epistemic success	The distance between agent's opinions and the truth value of the issue.	The number of agents who prefer the strongest research program.

Table 8.1: A comparison of the main features of the two agent-based models of scientific communities presented in Chapter 6 and Chapter 7.

Model	Result
Chapter 6	Agents' opinions stabilize quickly when investigating issues which are close to falsity or total truth, and slowly when investigating issues whose truth values are in the middle of the opinion space.
	Communities become more diverse when connections in the network increase below a certain threshold, and less diverse when connections increase above the threshold.
	Diverse communities are less epistemically successful.
Chapter 7	Non-biased communities are good at selecting the best alternative.
	Biased communities are less epistemically successful than non-biased ones.
	Biased communities tend to polarize less than non-biased ones.
	The negative impact of the bias on epistemic success can be mitigated by an equal number of agents supporting each alternative in the beginning.
	The negative impact of the bias on epistemic success can be partially mitigated by a mechanism to filter out weak arguments at the community level, which we interpret as "shared beliefs".

Table 8.2: A summary of the results of the models of Chapter 6 and Chapter 7.

Part III

Exploring Applications

Introduction to Part III

The models presented in Part II demonstrate that the outcome of controversies are influenced by the way arguments are shared within a population. This highlights the importance of gathering detailed information about the situations being modeled. Thus, agent-based modeling alone is not sufficient to study controversies, especially when it comes to exploring their broader socio-political dimensions, as they can involve not only scientists, but also a variety of actors such as journalists, industrials, policy-makers, activists... In order to assess how the abstract tools I had developed could represent the complexity of such debates, I dedicated the latter part of my PhD to study real-life controversies using abstract argumentation.

A close contribution is that of [Salliou and Thomopoulos \(2018\)](#), which use argumentation frameworks to investigate the arguments influencing people to adopt vegetarian diets. This work offers valuable insights about the content and structure of the debate about vegetarianism. For instance, the study reveals that health-related arguments hold a dominant position in the debate, significantly outweighing ethical or environmental arguments in terms of influence. Additionally, certain arguments with high centrality in the network are identified as key points of contention or support, which tend to shape the overall debate. The findings suggest that discussions around vegetarian diets may polarize around these health-related arguments. Thus, this formal analysis shows that abstract argumentation can provide a deeper understanding of how debates are structured and offers insights into the most influential arguments driving the discourse.

Following a similar approach, I modelled the argumentation frameworks corresponding to real-life controversies, exploring how abstract argumentation can improve our analysis of these debates. This work also provided valuable insights about how to model controversies in agent-based models, in line with discussions within the ABM community arguing that researchers using agent-based models should increase their use of empirical data to inform their simulations ([Edmonds and Moss, 2004](#)).

This approach was developed in two projects. The work performed on these two use cases led me to identify an aspect which had been mostly absent of my previous modelling work: the question of **argumentative strategies**.

In Chapter 9, I study the arguments surrounding a debate over an industrial project: the modernization of a sawmill in Saint-Agnan, located in the French department of Tarn. To build the argumentation framework for this controversy, I employed innovative techniques, such as co-constructing the framework during interviews with the actors. I explored how visualizing the arguments allowed a reflexive exercise from the participants, and the analysis of the argumentation graph revealed distinct strategies employed by both sides of the debate.

Chapter 10 focuses on a well-known socio-technical controversy: the debate on the origin and existence of climate change. I took part in a project to characterize the *scientific* arguments employed by climate skeptics, and we used abstract argumentation and qualitative analysis to identify specific features of contrarian discourse on climate: notably, self-contradiction and the misuse of scientific arguments.

Visualizing Argumentative Strategies in Controversies over Contested Projects

9.1 Introduction

The acronym GPII ("Grands Projets Inutiles et Imposés") is used in France to describe large infrastructure projects that face significant opposition from local communities (Grisoni, 2015). Debates surrounding these projects can sometimes escalate to a national level, as seen in the controversies over the A69 highway in southern France or the high-speed rail line between Lyon and Turin in the Alps. These debates, which intertwine environmental, political, and technical arguments, are clear examples of socio-technical controversies. Unlike purely scientific debates, they are not driven by discussions between scientists but involve various actors, including activists, local residents, government officials, and industry representatives. As a result, they represent a unique type of controversy with distinct structures and interactions between different groups. To my knowledge, these debates have yet to be modeled or explored through abstract argumentation.

This particular use case focuses on the debate between proponents and opponents of building a large industrial sawmill in Saint-Agnan, located in the French department of Tarn. Compared to more high-profile environmental disputes, such as the decades-old and ongoing Lyon-Turin rail project, this case is smaller in scale and more recent, making it more manageable for detailed analysis. However, I propose that the method developed here can be successfully applied to larger, more complex controversies.

Infrastructural disputes are inherently political and, when they escalate into conflicts between opposing camps, they involve a variety of communication strategies. Many of these strategies depend on the dissemination of arguments for or against the project.

The central research question guiding this study is: **Can abstract argumentation contribute to analyzing and shaping these argumentative strategies?** I explored abstract argumentation both as a research tool for gaining deeper insights into the controversy and as a practical tool that could be used by the participants themselves, helping them better understand the debate and refine their own argumentative strategies.

I conducted several interviews with members of the collective opposing the sawmill's construction. This allowed me to 1) create a representation of their arguments using abstract argumentation and 2) present this representation to them to gather their feedback. At the same time, I used various documents to build a representation of the arguments from the sawmill's supporters, including the company building the sawmill and certain local elected officials. I then presented this argumentation graph to the members of the collective to assess whether it could help them refine or reconsider their argumentative strategy.

To the best of my knowledge, I am the first:

- to apply abstract argumentation to the representation of a controversy about infrastructures.
- to use sociological interviews to inform the construction of an argumentation graph.
- to study the strategical use of abstract argumentation for a group of interest.

The first section of this chapter is devoted to a brief presentation of the context. I present my method in the second section, and the results in the third section. The fourth section discusses how these results relate to the literature in both formal argumentation and the sociology of controversies.

9.2 Context

9.2.1 The "mega-sawmill" project

SIAT, one of France's largest players in the forest industry, recently acquired a small sawmill in Saint-Agnan, located in the Tarn department, with plans to significantly expand and modernize the facility. The proposed expansion would increase the mill's capacity nearly fourfold, processing approximately 450,000 logs per year compared to the current 100,000. The new facility is intended to be a "mirror" of SIAT's parent site in Urmatt, Alsace. This site is one of the largest sawmills in France, and has been previously linked to environmental pollution concerns.

A group of local residents has organized in opposition to what they see as an outsized and potentially harmful project, which they refer to as the "mega-sawmill". While few dispute the need to modernize the facility or question SIAT's right to do so, they are concerned about the scale of the expansion. Their worries center on the potential strain on local wood and water resources, the risk of pollution affecting air and water quality, and the overall impact on their quality of life. Additionally,

they criticize what they see as a poorly managed democratic process, citing a lack of public consultation in the validation of a project which is partially funded by public money. These residents are calling for a more participatory approach in shaping the modernization plan, particularly regarding the size of the new structure.

Unlike more high-profile protests against industrial projects, such as the opposition to the A69 highway in the same department, this particular conflict has not garnered national attention. All involved actors are local. To date, the opponents have focused their efforts on researching the project, constructing counter-arguments, and informing the local community through public meetings. They also run a public Facebook page¹ and a blog², where they share articles and updates on their findings, aiming to raise awareness and engage with the local population.

9.3 Method

9.3.1 A tool to represent argumentation graphs

To evaluate whether abstract argumentation could influence perspectives in this debate, I needed a way to visually present argumentation graphs to participants. Unlike the graphs used in the agent-based models from Chapters 6 and 7, I needed to represent an argumentation graph containing real arguments. For this purpose, I developed a program that displays both the structure of the argumentation graph and the content of the arguments. The program was built using Dash³, a Python-based data visualization library. To allow for easy modification of the arguments and their relations, the program uses a simple spreadsheet as input, which contains the arguments and their connections. It then generates an interactive visualization of the graph that can be accessed locally. Figure 9.1 shows the tool's interface upon launch, where users can choose to display one of three graphs: the CON graph, the PRO graph, or a merged version of both.

9.3.2 Co-construction of the argumentation framework of the opponents

I conducted four interviews with two members of the collective, which will be called S. and E. in this chapter. S. and E. were already part of a victorious fight of local inhabitants against the construction of a mine in a nearby village. This experience led them to become active members of the collective "Brassac-Sur-Agout & Le Bez"

¹<https://www.facebook.com/people/Collectif-Brassac-sur-Agout-Le-Bez/100082323695257/>

²<https://wegreen.fr/group/informations-sur-le-projet-de-megascierie-de-saint-agnan-aussi-dit-de-brassac/stream>

³<https://pypi.org/project/dash/>

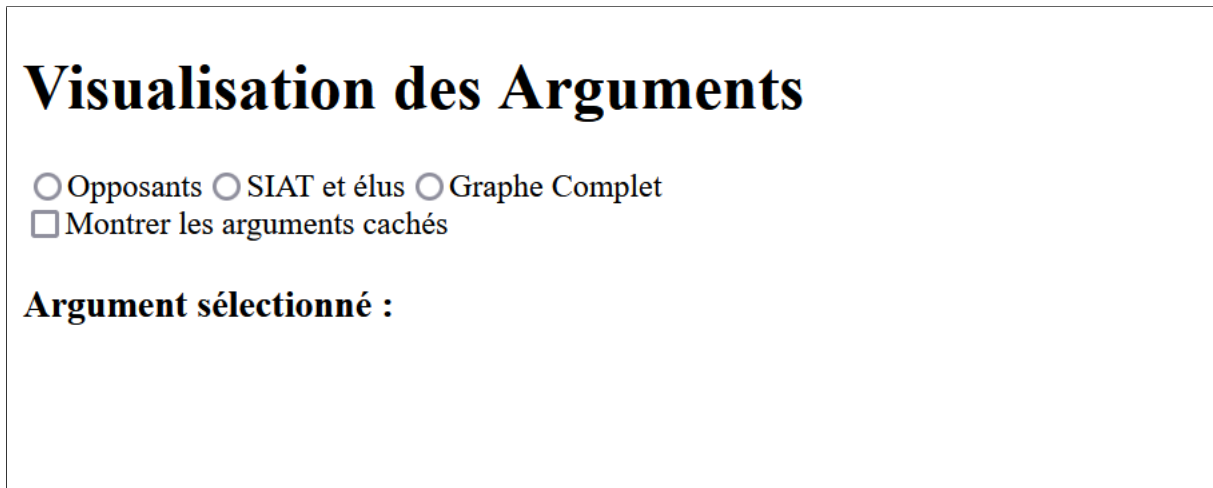


Figure 9.1: Screenshot of the visualisation tool. The user is prompted to select between the CON graph ("Opposants"), the PRO graph ("SIAT et élus") and a merge of both ("Graphe Complet"). "Argument Sélectionné" shows the details of the selected argument.

which protests the construction of the sawmill. S. and E. are executives who settled in the region about ten years ago.

The interviews were conducted as follows:

1. **First interview:** S. presented the global context, and I asked him to list all the arguments of the collective of opponents against the sawmill.
2. I built an argumentation graph using the arguments collected during the first interview.
3. **Second interview:** I showed S. the argumentation graph using the visualisation tool, and asked him whether he agreed on the relation between the arguments.
4. Using the input from S., I modified the graph, adding or removing some relations and arguments.
5. **Third interview:** I met E; and showed her the argumentation framework, collecting her inputs.
6. I realized a final session of modifications of the graph. I also added a few arguments obtained from the web comments under a local press article about the project, which had not been expressed by the two interviewees.
7. **Fourth interview:** I showed S. and E. the final graph, and we had a discussion about the potential use of this tool for the collective.

The argumentation framework obtained from these back-and-forth discussions will be denoted the CON graph.

9.3.3 Creating the argumentation framework of the supporters of the project

In parallel, I created a second graph which represents the arguments of the "pro" side. I used a written source: the slides of a meeting organized by the company SIAT and local elected officials to present the project to local residents. I manually extracted the arguments and relations from this document. This is a usual method to construct argumentation graphs: for instance, [Polberg and Hunter \(2018\)](#) and [Salliou and Thomopoulos \(2018\)](#) built argumentation frameworks by manually analyzing written texts. As the source document is very structured, it was comparatively easier to translate it into arguments. This graph will be denoted the PRO graph.

During the interviews, I also showed S. and E. the PRO graph, and collected their impressions. This was a way to study whether the representation of arguments using abstract argumentation gave them a better understanding of the strategies mobilized by the supporters of the project.

9.4 Results

9.4.1 The CON graph

Figure 9.2 presents the argumentation graph of the opponents to the projects.

The structure of the graph is similar to what I defined in my framework of Chapter 5: the graph is a bipolar, *issue-oriented* argumentation graph. It does not feature any cycle. A central argument or "issue" represents the main focus of the debate: the project of modernization of the sawmill, here denoted "Projet SIAT". Attack relations and support relations are distributed in a specific way: there is a ring of attackers of the central issue, outlined in bold on Figure 9.2. The rest of the argument are direct or indirect supporters of these attackers. Thus, all of the arguments are direct or indirect attackers of the issue.

The structure of the graph evolved throughout the interviews: during our first discussion, S. presented some of the arguments of the pro side, to explain the argumentative strategy of SIAT and the counter-arguments of the opponents. When reworking the graph, these arguments disappeared as they were not necessary anymore: rather, they were turned into attacking arguments. For instance, S. had identified that SIAT promoted an argument about the respect of regulation, "SIAT respects the environmental regulations", and told me that the group of opponents countered this argument with proofs that they did not respect the environmental rules in Alsace. In the final graph, I transformed this into the negation of the argument, namely "SIAT does not respect the environmental regulations", which allowed

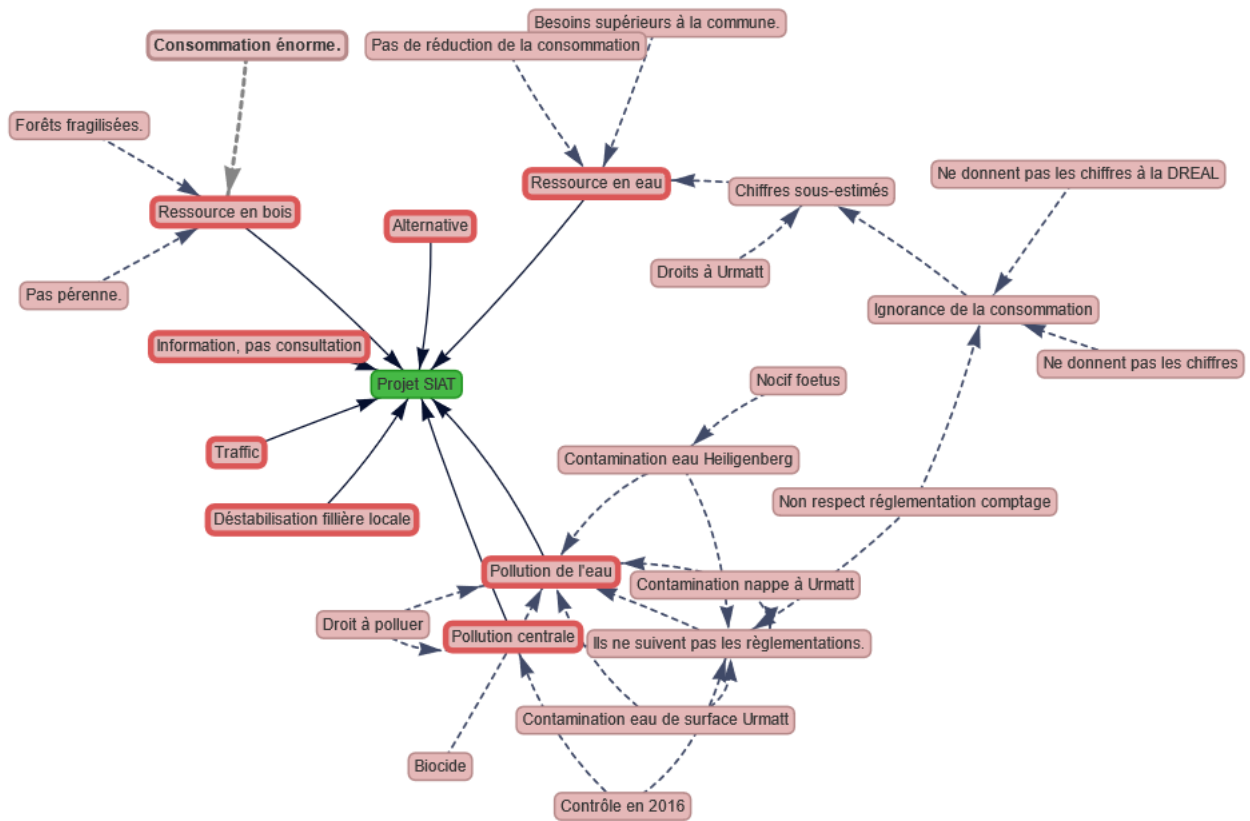


Figure 9.2: Representation of the CON graph, the argumentation graph constructed with the member of the collective of opponents. Dashed arrows represent support relations and full arrows represent attack relations. Arguments outlined in bold are direct attackers of the *issue*, in dark green.

for a much easier understanding of the graph. Thus, the exact shape of the graph depends on choices made by the modeler, for reasons of readability, symmetry, etc.

One section of the graph represents a rather complex group of arguments: those centered on the pollution of water and air by the project. Figure 9.3 clarifies the relations between these arguments. The first argument in support of the claims that the water and air of Saint-Agnan will be contaminated is the fact that environmental regulations do not prevent pollution, but rather allow it within a certain limit: thus, the company has a "right to pollute" below a certain threshold. In addition, opponents have found several examples of water and air pollution at levels above the legal limit at two SIAT sites in eastern France. These arguments support, respectively, claims that Saint-Agnan's water and air will also be contaminated, but they also support a more general argument by opponents, namely that SIAT has a history of non-compliance with pollution regulations. This is an important argument, as we will see below that a large part of the company's argumentative strategy is based on its assertion that it will comply with the regulations. Because of this structure, the example of air contamination in Urmatt indirectly supports the claim that the water in Saint-Agnan will be contaminated, and similarly, the water pollution examples support the likelihood of air contamination. From the point of view of abstract argumentation, it is a good strategy on the part of the opponents, who make the most of the arguments they have.

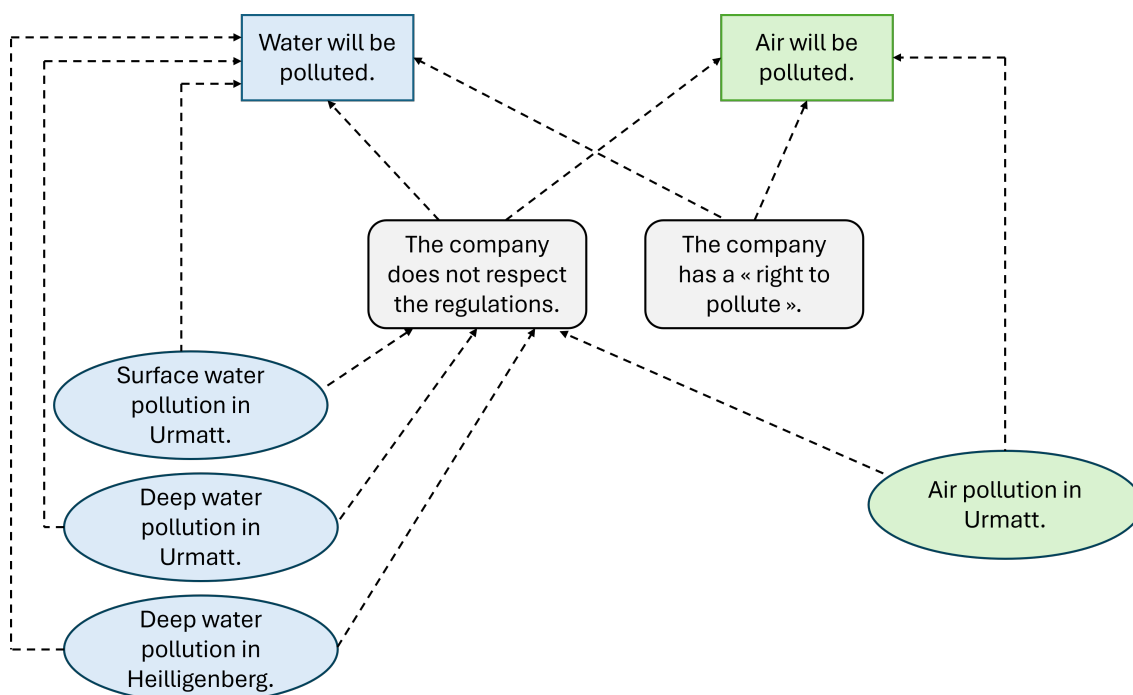


Figure 9.3: Diagram clarifying the structure of the water and air pollution arguments within the CON graph. Dashed arrows represent support relations.

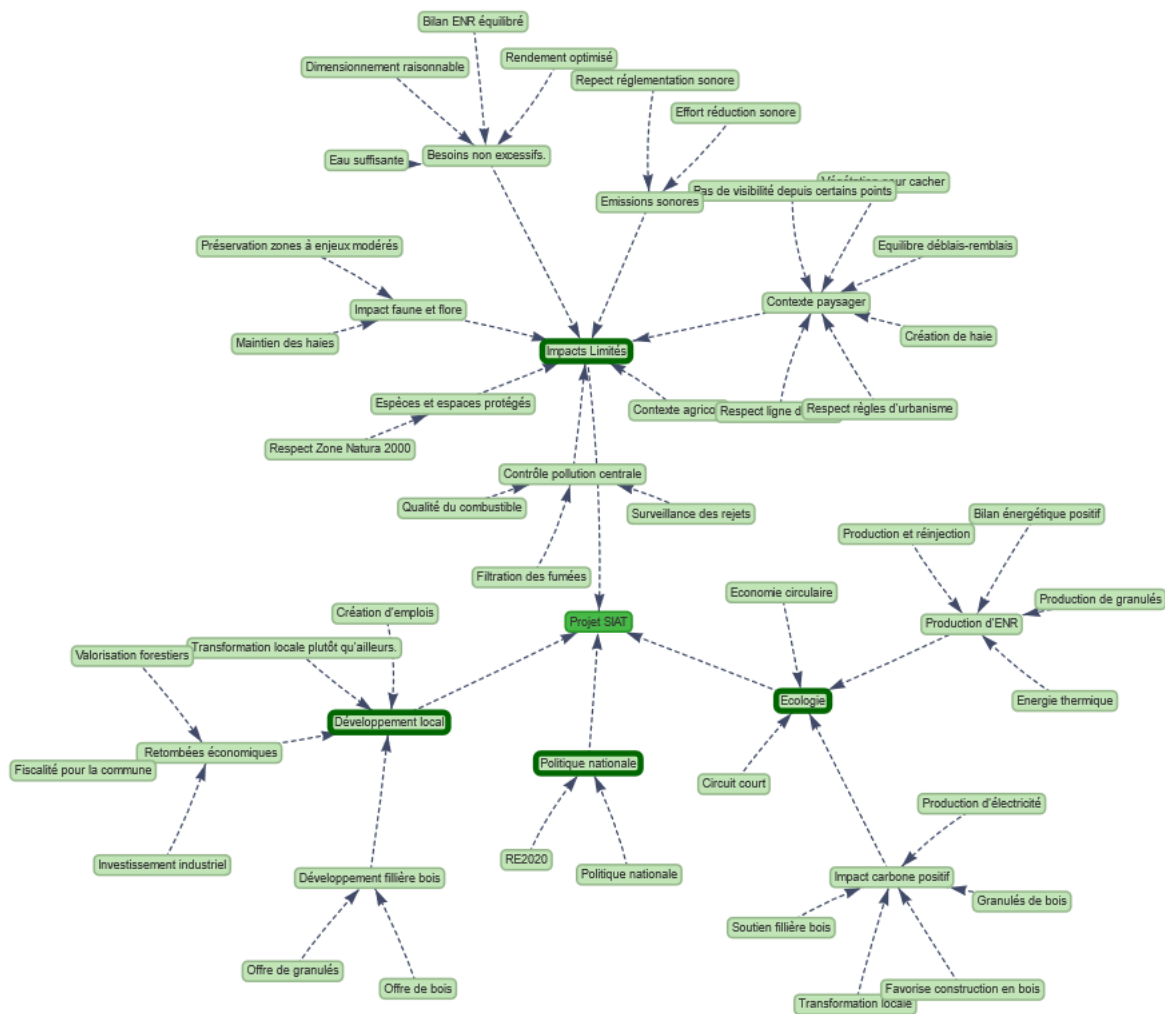


Figure 9.4: Representation of the PRO graph, the argumentation graph of the supporters of the project. Dashed arrows represent support relations. Arguments outlined in bold are direct supporters of the *issue*, in dark green.

9.4.2 The PRO graph

Figure 9.4 represents the graph of the supporter of the project. This graph is also an issue-oriented bipolar graph, with no cycles. We can note that it is composed only of support relations: understandably, the company SIAT did not provide any arguments against its own project in its public presentation to local people.

A big part of the arguments of the PRO side rests on the respect of existing laws and regulations, and on the fact that the project will not have a strong impact on the local population: this is what is under the argument "Impacts limités". Thus, we can see that the PRO graph is indirectly answering the objections of the CON graph, although these objections are not explicitly formulated. Another pillar of their argumentation is the economic impact of the project on the region:

argument "Développement local". Finally, the supporters of the project argue that it is beneficial for the environment: argument "Ecologie", claiming a positive carbon footprint due to the production of renewable energy.

9.4.3 Merging the graphs

The two graphs have one argument in common: the issue. The CON graph has a complex structure, while the PRO graph is a tree, each argument supporting exactly one argument. This difference may be due to the type of support used to collect the arguments: the written format, which was used to build the PRO graph, allows for a more structured discourse.

In order to merge the graphs, I represented all arguments and existing relations, and added some additional relations between them. The only relations that were added were attack relations from the arguments of the CON graph towards the arguments of the PRO graph. This represents the dynamics of the debate: SIAT presented its project first, and it was attacked afterwards by the opponents who found ways to rebut some of the company's arguments. In the future, it is possible that the company may issue some counter arguments to those of the opponents, which would further change the shape of this graph.

Figure 9.5 presents a visualisation of the merged graph.

This graph has not been reviewed by the participants: this will be done as future work. There are several choices which, in my opinion, are open to discussion: for instance, the argument stating that «SIAT does not respect regulations» could attack a great number of arguments of the PRO graph, which rests on the promise that SIAT will respect regulations, e.g. for endangered species, for urbanism, etc. However, since this argument mostly concerns the water and air pollution, I chose not to include these attack relations. Some researchers use argumentation graphs with weights on attack or support relation to denote their strength. If I was able to assign weights to attack or support relations, a solution could be to add an attack relation from this argument to all arguments concerning regulation, with a higher weight for the water and air pollution claims.

9.4.4 Informing argumentative strategies

The three graphs give us insights about the two groups' respective argumentative strategies.

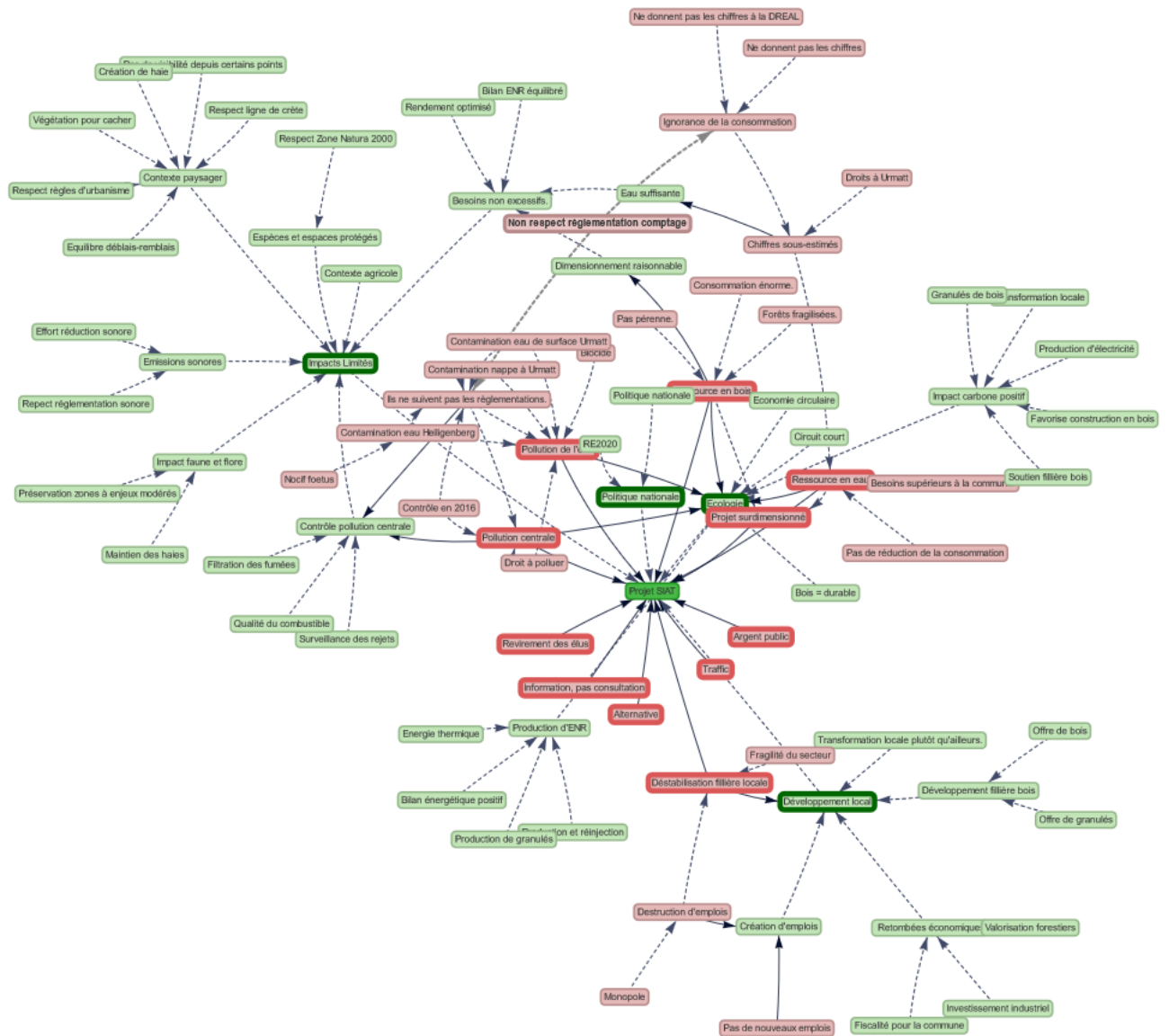


Figure 9.5: Representation of the merged graph, the argumentation graph combining those of the supporters (in green) and opponents (in red) of the project. Dashed arrows represent support relations, full arrows represent attack relations. Arguments outlined in bold are direct supporters or attackers of the *issue*.

Gish Gallop One thing that we can note is that the two graphs do not feature the same level of details: some of the arguments of the PRO graph are more trivial: for instance, one of them states "We will build a hedge" as a reason to support the project. Another argument states that they "will respect the protected area", which is required by the law: simply stating that something is not illegal is hardly a strong argument for a project. This is not to say that all the arguments of the PRO graph are weak, but that among its many arguments a certain proportion is trivial. The CON graph features less arguments, but most of them seem strong and many are backed by data obtained by the opponents.

This difference in the number and type of arguments could be a difference in the strategies used by the groups: the presentation made by SIAT was obviously a rather exhaustive document listing all the reasons for the project, however trivial. This looks like a debate strategy called "Gish Gallop": defined by [Scott \(1994\)](#), it consists in flooding the opponents with a great number of arguments with no regard for their strength. This strategy rests on the fact that the opponent will not have the time to counter all of these arguments and that the public will not have the time to accurately gauge their strength: in the end, the accumulation of trivial evidence gives the impression of a strong case.

This difference in the importance of the arguments motivates the use of *weights* to represent the intrinsic strength of the arguments. This ensures that a semantics would not treat, for instance, the arguments "We will build a hedge" and "The water was found to be contaminated 300 times above the legal limit" as having the same impact. Adding such weights and applying a semantics would be the next step of this project.

Local vs global The supporters of the project advance arguments about greenhouse gas emissions: for instance, that the sawmill project would have a positive carbon footprint by favoring the use of wood as a heating source. As we can see on the merged graph, this subject is completely ignored by the opponents, who focus on local environmental impact, and especially on impacts on the quality of life of local residents. When I formulated this remark to S. and E., they confirmed that this was a conscious strategy of the opponent group: they told me that "[local] people do not care about the harm done to biodiversity, they care whether their water will be poisoned". Thus, they focused on arguments most likely to sway the local population, such as the direct danger to their quality of life. However, this lack of consideration of the overall environmental context means that there are many arguments of the supporters of the project which are not challenged, in particular those claiming that the project is ecological. One may wonder whether convincing the local inhabitants is the best, or the only available strategy for the opponents. Different argumentative strategies may be relevant for different types of public: local inhabitants, elected officials, state representatives, and even national opinion, that the collective may want to convince if they hope to make their voice heard.

This tendency to focus on the quality of life and ignore the global context is characteristic of many local opposition groups and can be termed the "NIMBY" strategy. "NIMBY", which stands for "Not In My Backyard", is an acronym which designates people who are opposed not to the project itself, but to the fact that this project would be close to their homes ([Hermansson, 2007](#)). The problem for these groups is that such arguments are easily swayed by appeals to the common good of the type "We need this infrastructure, it has to be built somewhere.", and they can be accused of selfishness. As our use case illustrates, this leaves the supporters of the project in the position of promoting the global good, which is a definite argumentative advantage. When an opposition group goes beyond the NIMBY philosophy to oppose the project from a global perspective, it gains argumentative respectability, although this implies greater opposition and anti-systemic character. Take, for instance, how the opposition movement against the Notre Dame des Landes airport in France adopted the slogan "Neither here, nor elsewhere" to signify its opposition to *any* airport project, which was instrumental in their victory ([Troupe, 2018](#)).

Similarly, the question of local economic impact is one of the pillars of the argumentation of the supporters of the project. The merged graph shows us that this aspect was not developed by the opponents: they only have one attacking argument, which states that the modernization of the sawmill will destabilize the local wood industry. Throughout our discussions, we identified this as a subject that they could develop further: for instance, by stating that smaller, local sawmill create more employment than large ones and are disappearing in France ([d'Allens, 2019](#)).

Thus, having access to the abstract representation of the arguments promoted a better understanding of both groups' strategies. This tool allowed me and the participants to identify some ways to improve their argumentative strategies. It also highlighted the importance of weights to describe the intrinsic strength of arguments. These results favor the idea that abstract argumentation can be a useful tool to improve argumentative strategies and allow for the better understanding of a debate.

9.5 Related Works

Using abstract argumentation to study a public debate, this work is part of at least two distinct literatures: on the one hand, empirical studies of abstract argumentation, which aim to verify the applicability of abstract argumentation approaches in the real world, and, on the other hand, the study of controversies: a field of sociology that describes and analyzes argumentative discussions in the public sphere.

9.5.1 Empirical studies of abstract argumentation

The empirical studies of abstract argumentation presented in Chapter 2 investigate whether certain principles of abstract argumentation semantics correspond to the way humans reason (Rahwan et al., 2010; Cramer and Guillaume, 2018, 2019). They use methodologies inspired by psychology and cognitive science, usually recruiting anonymous participants who perform predefined tasks. This contrasts with my methodology of conducting long and detailed interviews, which is closer to that used in sociology. Despite these methodological differences, several empirical works on abstract argumentation offer interesting points of comparison.

For example, the study by Polberg and Hunter (2018) supports the use of bipolar graphs as a more accurate way to represent how people reason. This finding inspired me to incorporate bipolar graphs into my own research. The authors also observed that when participants were asked to determine the relationships between arguments, the resulting graphs varied significantly, often differing from the relationships predicted by the authors. This suggests that there isn't always a single correct graph corresponding to a debate. In my own work, when we were co-constructing the graphs, I similarly noticed that some decisions were somewhat arbitrary—for instance, I chose to merge certain arguments to enhance readability or simplify the graph. This observation implies that multiple representations of the same debate can be equally valid. Polberg and Hunter (2018) propose a constellation probabilistic approach to account for uncertainty in the structure of the graph. While uncertainty may not be the perfect fit for my context, as it presupposes a single objective truth, this type of approach might still be applied in a fruitful way to accommodate the variability in how debates can be represented.

Vesic et al. (2022) conduct experiments to test, among other things, whether humans who are provided with the graphic representation of a debate draw conclusions that are more aligned with the predictions of abstract argumentation. This research question is quite close to mine, as I am testing whether abstract argumentation (and especially its visual representation) help people to create better argumentative strategies. In both cases, the role of the visual representation of the argumentation graph is questioned. The findings of Vesic et al. (2022) support the idea that graphical representations make participants more likely to adhere to the abstract principles of gradual semantics. The authors also highlight that the acceptability of arguments isn't solely determined by the structure of the graph—participants rely on their prior knowledge to judge which arguments are acceptable. To account for this, the authors propose using weighted argumentation graphs. I reached a similar conclusion, observing that it was impossible to disregard the varying strength of the different arguments I collected.

Thus, the work presented in this chapter corroborates two findings from the empirical abstract argumentation literature: the relations of attack and support that characterizes a debate is not unique (Polberg and Hunter, 2018), and weighted argumentation graphs are more likely to represent how people reason about arguments

(Vesic et al., 2022).

9.5.2 The study of controversies

Latour and Woolgar (1979a) famously pioneered the *science studies*, the anthropology of science and technology. They especially explore scientific controversies, which are debates between proponents of different scientific theories. Lemieux (2007) proposes a definition of a controversy, which includes 1) A triadic structure, two opposing camps and a real or virtual "public", 2) A form of egalitarian treatment of both sides (although often very asymmetrical) in their ability to get arguments across to the public and 3) A rather closed environment, of "specialist" peers sheltered from the general public, who are the only judge.

We can see that our use case - the debate between proponents and opponents to a construction project - partially fits this definition of a controversy. The structure is triadic, with both camps trying to convince a public: often, the local inhabitants; sometimes, national opinion or state officials. There is a form of egalitarian treatment, at least in the capacity to get arguments across to the public: for instance, in the case that I studied, both parties were able to organise public meetings with local people; in reality though, there is often a large asymmetry in the degree of institutional and media support that the proponents of a project receive, compared to the opponents. Where this use case diverges clearly from a controversy in the sense of Lemieux (2007), is that the debate is not confined to a public of "peers": rather, it is characterized by a diversity of actors: the opponents can be from all backgrounds, and the proponents are usually members of companies or state institutions. The final decision is not taken by the public itself, either national or local, but by state institutions. Other controversies which do not feature a sheltered public of peer have been defined as **socio-technical controversies**; a famous example being the debate about the use of electric cars (Catellani and Lits, 2024). In a sense, all ecological controversies could be considered socio-technical controversies, since they mobilize scientific arguments about ecological consequences: CO2 emissions, impact on biodiversity, energy consumption, and so on. As such, the findings from the sociology of controversies can be applied to our case.

In particular, Akrich et al. (2006) defended the idea that, unlike what is classically analyzed, controversies do not reveal existing oppositions, with defined groups and pre-built arguments, but rather that they are social facts that *create* a new reality: that groups of partisans and arguments are developed and consolidated throughout the conflict. This idea resonates quite interestingly with my current work: indeed, as they built and reflected on the argumentation graph, the participants were able to think of new arguments and strategies. Some of these strategies imply a shift in their perspective, such as the one which consists in opposing the project from a more global socio-economic perspective. Thus, this experience supports the idea of Latour, Callon and Akrich that indeed, the process of the

controversy itself creates the different positions of the actors.

Šešelja and Straßer (2013) were pioneers in representing scientific controversies using abstract argumentation. They notably explored a controversy from the 1920s involving Alfred Wegener and the proponents of continental drift theory, who were competing with other scientists to explain geological phenomena, such as the presence of similar fossils on different continents. This work influenced our approach to modeling research programs in the agent-based model presented in Chapter 7. Inspired by the analysis of Šešelja and Straßer (2013), we represented scientific theories as collections of "central arguments." These central arguments form the core support for each theory: for instance, the observation that the continents resemble pieces of a jigsaw puzzle was a key argument in favor of continental drift. They serve as the foundation for independent sub-debates.

In the debate surrounding the Saint-Agnan sawmill, we can observe that the main points of contention or support, highlighted in bold within the graphs, could also be considered central arguments. However, unlike in the modelling proposed in Chapter 7, these arguments do not appear to anchor independent sub-debates. For example, in the diagram of Figure 9.3, the arguments supporting "Air pollution" and "Water pollution" clearly interact. This could be because scientific debates are more likely to foster compartmentalized discussions around specific pieces of evidence, whereas debates like the sawmill controversy are broader and more politically charged. Alternatively, it is possible that an alternative framing of the debate could reveal a structure that better aligns with the central argument approach, though such a structure did not emerge from my current method.

It seems that the study of controversies is a fruitful field of application for abstract argumentation. This creates interesting links between abstract argumentation and the Actor-Network Theory (ANT) developed by scholars like Bruno Latour, Michel Callon, and John Law in the late 20th century (Latour, 1999). ANT posits that society is made up of networks of both human and non-human actors (or "actants"). These actors, which can include people, organizations, technologies, and even concepts, interact and influence one another to form networks. Among the specificity of Actor-Network Theory is the idea of symmetry: ANT treats all actors (human and non-human) equally in their ability to influence the network. For example, a technology or a concept can have as much impact as a person. ANT is often used in fields like sociology, anthropology, science and technology studies to explore how power, knowledge, and social order are produced and maintained within complex systems. It typically requires mapping out the relationships and connections among the actors. Pacôme Delva and Nestor Herran, two researchers who offer a course on modeling controversies using Actor-Network Theory (ANT) for first-year university students, have adapted my visualization tool to model the debate studied in their class: the discussion around electric cars. They modified the traditional graph format to align with ANT principles, incorporating not just arguments but all actors involved in the debate—such as the government, media, arguments, concepts, and policies. In this model, any actor can support or attack

another; for instance, the government might support certain policies. This approach creatively applies ANT by treating all actors as potential arguments, resulting in a new form of hybrid argumentation graph. Naturally, further investigation is needed to determine whether the principles of abstract argumentation can be applied to such a graph. For example, can an entity like "the government" be "defeated" in an argumentative sense? What would an acceptability score mean in this context? These questions open up intriguing possibilities for the interplay between abstract argumentation and Actor-Network Theory.

9.6 Discussion and Conclusion

This work has yielded valuable insights on multiple levels. First, it assesses the suitability of the graph structures used in my theoretical research for representing real-world debates. Additionally, it evaluates a tool designed for visualizing abstract argumentation graphs and a methodology for conducting extensive interviews with participants. Finally, this research enhances our understanding of argumentative strategies.

Insights on the structure of the graphs This exercise provided valuable insights into which graph structure is best suited for modeling real debates. I opted to use issue-based bipolar graphs without cycles, and found this approach effective in representing the debate, supporting their application in theoretical contexts. However, one of my findings is that there isn't a single, unique argumentation graph that can describe a debate: this suggests that other structures could also be suitable. For example, although I did not end up with a graph featuring central arguments with independent sub-debates, as in [Borg et al. \(2017a\)](#), it's possible that, had I chosen this structure, the arguments could have been represented accordingly. For instance, a central argument like "The project will cause pollution" could have encompassed all the arguments in Figure 9.3. This underscores the importance of developing methods to represent uncertainty within graph structures, such as the constellation approach proposed by [Polberg and Hunter \(2018\)](#). Additionally, consistent with [Vesic et al. \(2022\)](#), my work highlights the necessity of assigning *weights* to arguments to accurately reflect a debate. It also suggests that weighting attack and support relations may be beneficial as well.

Insights on the tool and methodology Although we engaged in many insightful discussions about the arguments and the broader context of the project, it was challenging to keep the participants focused on evaluating and refining the relationships between arguments. Despite my efforts to make the graph as clear as possible, reading and correcting these relationships remained a tedious task, even for participants with a background in technical tools (e.g., S. who previously worked

as a computer scientist). Moreover, conveying a complex group of interrelated arguments, such as the one shown in Figure 9.3, is difficult using a network graphic library like the visualization tool. Achieving clarity requires manual, "hand-crafted" adjustments. Given these observations, I believe this tool is not yet ready to be used in isolation. Instead, it should be supplemented with additional graphs and analyses, like those I conducted earlier. As it is, this tool serves as a visual aid, designed to complement the analysis of someone knowledgeable in argumentation graphs. A possible direction of work could be to develop a more ergonomic tool that could be used by people who have no prior knowledge of abstract argumentation.

Insights on argumentative strategies Analyzing the argumentation graphs from both sides of the debate reveals distinct argumentative strategies. The PRO graph features a strategy similar to a Gish Gallop, characterized by an accumulation of weak arguments. In contrast, the CON graph exemplifies a NIMBY (Not In My Backyard) approach, concentrating solely on local issues with a strong emphasis on the residents' quality of life. These graphs enable participants to reflect on their strategies, allowing them to identify weaknesses in their own approaches. As a future development, one could envision an abstract argumentation-based intuitive tool that empowers local communities to develop effective argumentative strategies while also contributing to the theoretical understanding of argumentative dynamics through the accumulation of data and case studies.

CHAPTER 10

"Casting Doubt on Consensus": The Use of Scientific Arguments to Promote Climate Contrarianism

10.1 Introduction

In 2024, only about 78% of Americans reported that they believed in climate change, while a mere 54% stated that they believe it is caused by human activities ([AP-NORC, 2024](#)). Yet, the consensus of the scientific community on the reality of Anthropogenic Global Warming (AGW)—that is, the existence of climate change caused by human activities—has been established for decades. This divergence between public opinion and scientific consensus cannot simply be attributed to a lack of rationality: research has shown that climate denial is deliberately *engineered*.

Naomi Oreskes and Erik Conway describe the infamous "tobacco strategy" in their seminal book, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming* ([Oreskes and Conway, 2011](#)). This strategy was originally used by the tobacco industry to successfully create the *illusion* of a scientific controversy regarding the harmful effects of tobacco, long after the scientific consensus was established. To do this, they employed tactics such as producing biased or ambiguous science and manipulating mainstream media coverage. The goal of the strategy was to sow doubt, making it appear as though there was a legitimate scientific debate between two opposing camps, even when the scientific community overwhelmingly agreed on the harmful effects of tobacco. [Oreskes and Conway \(2011\)](#) demonstrate in their book that the controversy around global warming is another example of the "tobacco strategy," where similar tactics have been employed to foster doubt about the reality of AGW. Despite the overwhelming majority of climate scientists agreeing that climate change is real and driven by human activities, they have shown that actors, including industries and conservative think tanks, have successfully perpetuated the narrative of a divided scientific community for years—a strategy that continues to this day.

The authors even reveal that some of the same scientists who worked with the tobacco industry were later recruited by oil and gas companies, and that conservative networks supported these misinformation campaigns. Illustrating this, Matthew B. Crawford, a prominent philosopher who once worked at the conservative think tank George Marshall Institute, testified that "the trappings of scholarship were used to put a scientific cover on positions arrived at otherwise. [...] Part of my job consisted

of making arguments about global warming that just happened to coincide with the positions taken by the oil companies that funded the think tank" (Crawford, 2009).

Weatherall et al. (2020), in their article entitled "*How to Beat Science and Influence People: Policy Makers and Propaganda in Epistemic Networks*", used their agent-based model of scientific debates to simulate the tobacco strategy. They model two aspects of the strategy identified by (Oreskes and Conway, 2011): the selection and promotion of independent research which happens to support the industry's position, and the funding of additional research whose results are selectively published. They show that the strategy is extremely effective, as rational actors form a wrong opinion based on the evidence that is presented to them. Their model highlights the role of the diffusion of scientific arguments to sow doubt. The strength of the demonstration of Weatherall et al. (2020) is to show that there is no need to produce bad science to use the tobacco strategy: in the model, all the scientists work in good faith. The manipulation lies in the *circulation* of these arguments: by showing a biased sample of evidence to policymakers, miscellaneous actors or medias create the illusion of a balanced debate instead of an overwhelming consensus. This result highlights how agent-based models, and in particular social epistemology, can be powerful to represent these mechanisms of misinformation.

The tobacco strategy amounts to **using science to defeat science**. In order to understand and model socio-technical controversies using abstract argumentation, we need to know more about the way science is leveraged *as an argument* by climate contrarians. Exploring this controversy is extremely relevant to understand the argumentative dynamics surrounding cases of misinformation in similar socio-technical controversies.

The work presented in this chapter is part of a project studying the scientific literature mobilized by climate contrarians, which was launched by Maxime Woringer from the association *Transitions*¹. Joining the project, I used my knowledge of abstract argumentation to provide a specific argumentative perspective into the issue.

Related works

Several studies have explored the rethoric of climate contrarians² using numerical tools.

Recently, Chavalarias et al. (2023) investigated the climate contrarian sphere on Twitter, shedding light on the strategies employed by climate skeptics. Through a

¹<https://asso-transitions.org/presentation/>

²The term "climate skepticism" was originally used to describe those who raise doubts or questions about aspects of climate science. In this work, we favor the term **climate contrarianism**, which has been proposed to highlight the organized and persistent opposition to climate science.

temporal analysis of the volume and content of posts, they revealed a significant *intensification* of climate skeptic activity in 2022, particularly marked in France by the creation of a large contrarian community. The study highlighted key differences between climate experts from the Intergovernmental Panel on Climate Change (IPCC) and contrarian accounts: while scientists focus their communication on their areas of expertise, contrarians tend to comment on a wide array of topics, projecting a false sense of expertise. The behavior of these contrarian accounts also pointed to the presence of numerous bots and suggested coordinated actions among them. This study emphasizes that, despite decades of work by the climate science community, climate misinformation remains a pressing issue. It underscores the importance of examining the argumentative tactics employed by climate skeptics, as their influence very much continues to shape our political landscape.

The centrality of scientific arguments in climate contrarian media is illustrated by the work of [Sharman \(2014\)](#), who analysed the type of content produced by climate sceptics sources. Using social network analysis, the author examines the climate skeptic blogosphere from March to April of 2012, classifying 171 blog posts as either science- or policy-oriented. She finds that blogs which are "central" in the blogosphere network tended to focus on discussions of science, while peripheral blogs tended to emphasize discussions about policy.

A common tactic of climate contrarians, whether they communicate via blogs, social networks, or think-tank websites, is to produce large lists of scientific studies which refute the consensus on anthropogenic climate change -or so they claim. For instance, in 2011, the Global Warming Policy Foundation -a think tank known for promoting climate denial and mistrust over climate policies - released an article with the headline "*900+ Peer-Reviewed Papers Supporting Skepticism Of 'Man-Made' Global Warming (AGW) Alarm*", which contained a list of 900 scientific papers which (according to the Foundation) supported a skeptic view of AGW³. The British climate news media *Carbon Brief* analysed this claim, studying the characteristics of these 900 papers. They found that, of the ten most prolific authors on the list, nine had links to organisations funded by ExxonMobil, the powerful oil and gas company, that 15% of the papers from the list had been published in a minor journal known for promoting skeptics views, and, finally and most surprisingly, that a large number of authors whose article was cited did not agree to feature on the list and refuted the interpretation of their work as promoting skepticism about climate change ([Carbon Brief, 2011](#)).

Since the 2010s, there has been a rise in studies applying bibliometrics—the use of statistical methods to analyze bibliographic data—to characterize the scientific controversy surrounding climate change ([Jankó et al., 2014, 2017](#); [Marx et al., 2017](#); [Fu and Waltman, 2022](#)). Notably, [Jankó et al. \(2014\)](#) examine the literature cited in the *Climate Change Reconsidered* report by the Heartland Institute, a well-known

³The original article has been taken down from the Global Warming Policy Foundation's website. The think tank now reports investigating the negative impact of climate change policies, and does not seem to promote skepticism about climate change anymore.

conservative think tank, and compare it to the science referenced in the Intergovernmental Panel on Climate Change (IPCC) reports. They find that the two bodies of science do not differ significantly in terms of the types of journals cited, and also provide an interesting comparison of the rhetoric used in both reports. However, their study lacks a more detailed investigation of the articles and their authors, which could reveal whether the science referenced in the *Climate Change Reconsidered* report fits the profile described by [Carbon Brief \(2011\)](#): for instance, whether the cited authors had potential conflicts of interest or whether they agreed with their work being used in the report.

In their article *"Computer-assisted classification of contrarian claims about climate change"*, [Coan et al. \(2021\)](#) conduct a thorough classification of the claims made by climate skeptics. They collect data from conservative think tanks (CTTs) and blogs, central actors in the climate change denial movement. First, the authors develop a comprehensive taxonomy of climate skeptic claims, and then train a deep learning model to classify them. With this tool, they construct a detailed historical account of climate change denial spanning two decades. Their study reveals a correlation between funding from dark money sources, the fossil fuel industry, and conservative donors, and the promotion of specific contrarian claims by CTTs.

The work of [Coan et al. \(2021\)](#) is highly relevant from an argumentation perspective, as their identification of claims provides a comprehensive taxonomy of climate contrarians' arguments. However, to my knowledge, no formal argumentation study has yet been conducted on the controversy surrounding anthropogenic global warming (AGW). Furthermore, I am not aware of any abstract argumentation research that focuses on the argumentative dynamics of the "tobacco strategy" or, more broadly, on techniques of misinformation. This research aims to bridge these gaps in the literature by employing both abstract argumentation and quantitative analysis to **characterize the scientific arguments of climate skeptics**.

We aim to answer the following questions:

- Can we explicitate the scientific arguments of climate contrarians using abstract argumentation?
- What are the characteristics of the scientific articles contrarians mobilize to support their claims?

For this, we propose to use three distinct and complementary methods. First, we leveraged the work of [Coan et al. \(2021\)](#) to construct an argumentation framework representing the skeptics' position. Then, we performed a qualitative analysis of the scientific content produced by prolific climate skeptic blog. Finally, we began building a database of the articles cited by these climate skeptic blogs, with the aim to perform a bibliometric analysis of the body of science mobilized by climate skeptics, and quantify the level of scientific support for each argument. This chapter presents the first two approaches in the two following sections. The construction of

the database, which is a work in progress, will not be detailed in this manuscript. However, an in-depth discussion about the relevance of this database is provided in the "Discussion and future work" section.

10.2 An Argumentation Framework of Climate Skepticism

Coan et al. (2021) have conducted an in-depth analysis of hundreds of thousands of articles from Conservative Think Tanks (CTTs) and blogs of climate skeptics. They extract five main claims promoted by climate skeptics:

1. Global Warming isn't happening - « it's not real ».
2. Human greenhouse gases are not causing Global Warming - « it's not us ».
3. The impacts of climate change are not harmful - « it's not bad ».
4. The proposed solutions won't work.
5. The science of climate and the climate movement are unreliable.

The authors also identify "sub-claims" that support the larger claims. For example, the claim "Global warming is not real" is backed by sub-claims like "The ice isn't melting," "The weather is cold," or "We are heading into an ice age." These sub-claims, in turn, can be further supported by more specific assertions: for instance, "The ice isn't melting" is reinforced by claims such as "Antarctica isn't melting" and "Glaciers are not vanishing." Figure 10.1 presents the taxonomy developed by Coan et al. (2021) mapping out this hierarchical structure.

What is remarkable about this taxonomy is that it mirrors the structure of an argumentation graph. The claims and sub-claims identified by Coan et al. (2021) form a hierarchical arrangement, where arguments support each other in a structure that closely resembles issue-oriented graphs. Additionally, the division into several trees with central roots (i.e., the main claims) is strikingly similar to the framework described in Chapter 7, where, drawing from the work of Šešelja and Straßer (2013), we modelled scientific theories as a collection of central arguments supported and attacked by subordinate ones. Interestingly, Coan et al. (2021) make no mention of abstract argumentation, and it is likely the authors are unfamiliar with it. This spontaneous resemblance further supports the relevance of abstract argumentation frameworks—particularly the hierarchical structures outlined in my work and that of Šešelja and Straßer (2013)—for representing and analyzing controversies.

The only element missing from this taxonomy, when viewed through the lens of abstract argumentation, are the **attack relations** between arguments. Indeed,

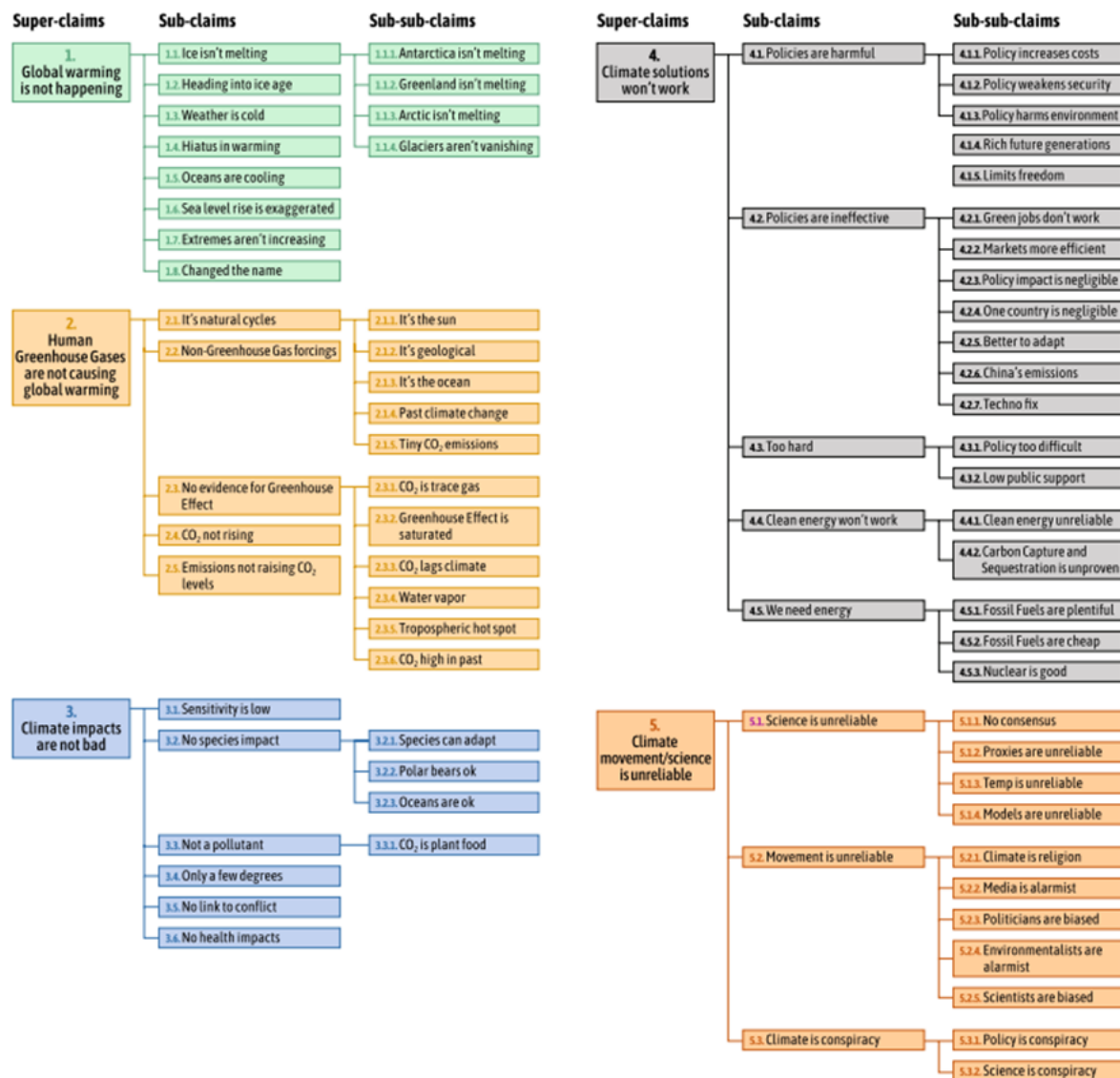


Figure 10.1: Taxonomy of contrarian claims of Coan et al. (2021)

several arguments in the trees for claim 2, "It's not us", and claim 3, "It's not bad", implicitly accept the reality of global warming as a premise. For example, sub-claim 3.4 asserts that the warming will only be a few degrees, which directly contradicts the claim that there is no warming or that we are entering an ice age (sub-claim 1.2). In Figure 10.2, I modified the taxonomy from Coan et al. (2021) to form an explicit argumentation framework. I introduced attack relations to represent contradictions between these claims, indicated by red arrows in the figure.

The specific structure of the attacks within the framework warrants further discussion. As demonstrated in the previous chapter, determining the exact relationships between arguments is not always straightforward. For example, whether the claim "Human greenhouse gases are not causing global warming" implicitly assumes the existence of a global warming is open to debate. However, one thing is clear: these arguments display internal contradictions. From an abstract argumentation perspective, this indicates a significant flaw in the set of climate contrarian arguments—it is not *conflict-free*, meaning it does not form an *extension* of acceptable arguments. Extensions represent coherent sets of arguments that a rational agent can assert (see Chapter 2). In this regard, abstract argumentation enables us to conclude that **the skeptics' position is not rational**, as it inherently contradicts itself. Moreover, if we compare it to Šešelja and Straßer (2013)'s characterization, we can say that the difference between the climate skeptic position and a legitimate scientific theory is this self contradiction.

A counterargument to this claim could be that the arguments presented by climate contrarians represent a collection of distinct positions. One could argue, for instance, that it is possible to assert all the arguments in the tree of claim 4, "Climate solutions won't work," without being irrational, as these arguments are not inherently contradictory. If this reflected the reality of climate skeptic discourse, the contradictory claims would arise from different individuals, specialized blogs, or websites. For example, some websites might deny the existence of climate change, while others might argue that climate change is not harmful. In such a case, we could consider multiple skeptics' positions. However, this is not representative of how climate skeptics typically assert their arguments. Take, for example, the Global Warming Policy Foundation, whose stated goal is to investigate the negative impact of climate change policies (GWPF, 2024). This organization might now be considered devoted to promoting claim 4 exclusively. Yet, this specialization is relatively recent: in 2011, according to Carbon Brief, the Foundation was sharing articles that questioned the reality of global warming, endorsing claims 1 and 2. Moreover, as we will see in the following section, our qualitative analysis of scientific arguments shared by climate contrarian blogs reveals that it is not uncommon for these media sources to assert contradictory claims. Therefore, in line with the findings of (Coan et al., 2021), we consider these arguments as forming a single position that is being actively promoted by climate contrarians. Results from the quantitative analysis of these articles may further refine this hypothesis.

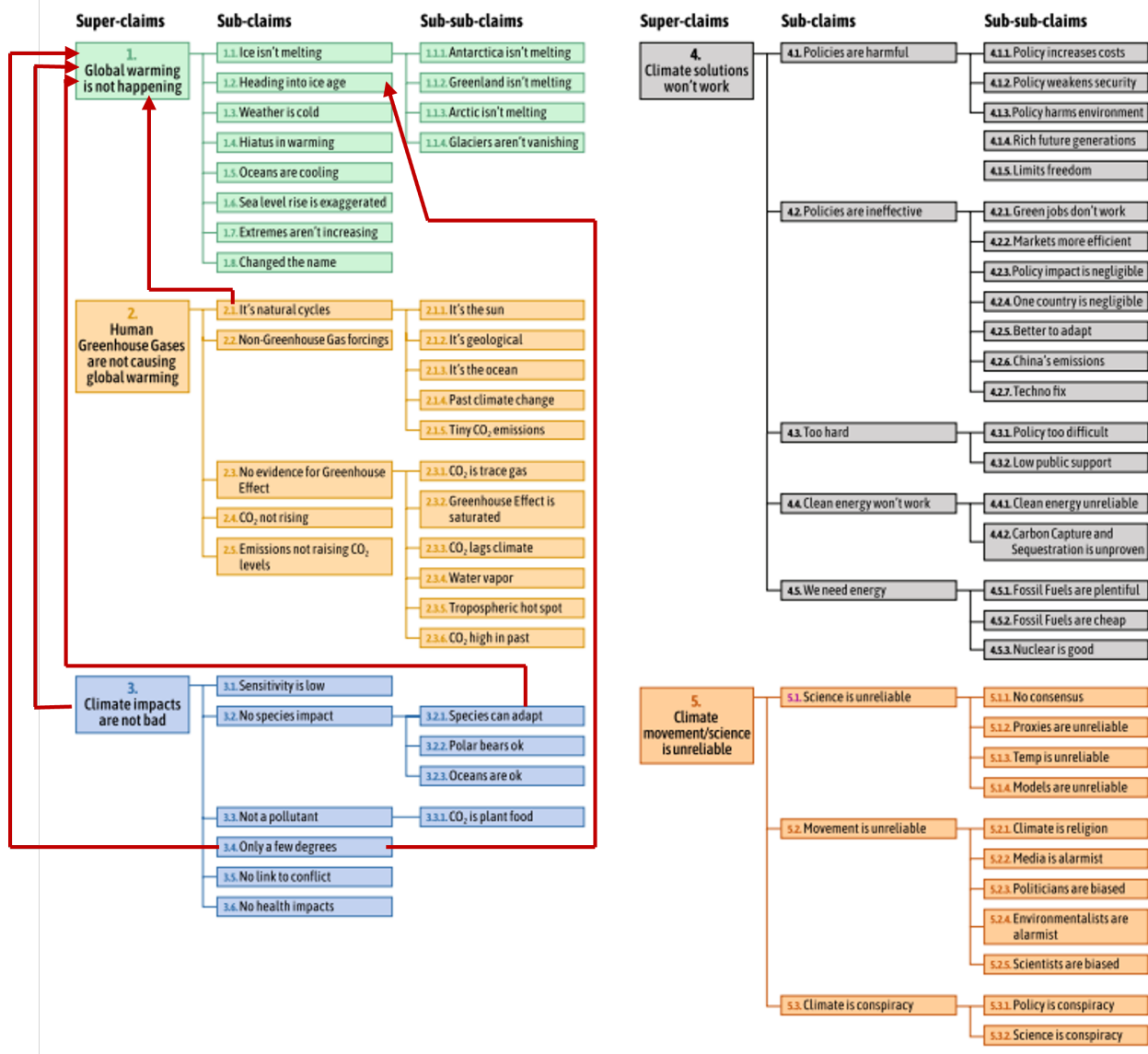


Figure 10.2: Abstract argumentation framework of the climate contrarian arguments identified by Coan et al. (2021). Red arrows represent attack relations, and the links between claims represent support relations going from sub-sub-claims towards sub-claims and from sub-claims towards claims.

Returning to our investigation of the scientific arguments used by climate contrarians, the analysis of this argumentation framework brings up important questions: which of these claims are, according to contrarians, supported by science? And to what extent can we quantify and qualify this scientific support? Addressing these questions requires further exploration. In the following section, we outline an approach that begins with a qualitative analysis of the articles cited by climate contrarian blogs, followed by a quantitative assessment.

10.3 Qualitative Study of Climate Skeptics Blogs

When they analyzed 900 scientific articles cited by climate contrarians, [Carbon Brief \(2011\)](#) highlighted the presence of three types of articles whose relevance was questionable:

1. Conflict of interest: articles whose authors are known to have received fundings from industrial actors interested in promoting climate skepticism (typically, oil and gas companies such as ExxonMobil).
2. Dubious science: articles which were not peer-reviewed, or published in prominent skeptics journals. For instance, the most cited journal on the list was suspected of questionable practices and co-edited by the director of the Global Warming Policy Foundation.
3. Misused articles: several prominent scientists whose work appeared on the list disagreed with the claim that their research supported skepticism about anthropogenic global warming. Some had previously requested, unsuccessfully, to have their work removed from similar lists.

Another category of papers included in the climate-skeptic corpus consists of legitimate scientific studies whose findings do not necessarily align with anthropogenic global warming. For example, these might be articles documenting localized decreases in temperature. The IPCC's methodology consists of assessing these results alongside numerous other studies reporting rising temperatures in various areas and across different time periods: this is how they conclude that global warming is occurring on average, without dismissing the existence of local exceptions to this trend. However, when these "contradictory" results are selectively highlighted and disseminated to suggest that global warming is not real, it represents a manipulation of legitimate science. This is the specific mechanism that was modeled by ([Weatherall et al., 2020](#)).

10.3.1 Description

Blogs play a crucial role in the dissemination of climate sceptic content ([Sharman, 2014](#)). We focused our qualitative analysis on a very prolific climate-skeptic website called *NoTricksZone* ([No Tricks Zone, 2024](#)). This blog is composed of dozens of pages listing scientific articles supporting various contrarian claims. These articles are grouped by themes -for instance, a page is name "Natural Oceanic Oscillations", another is titled "100+ Papers – Sun Drives Climate"- or by year, with archives dating back from 2014.

The pages grouping article by themes correspond to specific claims identified by [Coan et al. \(2021\)](#), which highlights the relevance of their classification. The blog indiscriminately asserts many different skeptics arguments, including contradictory ones. For instance, in the blog's home page, we can see that a link named "Cooling/temperature", which corresponds to the claim that the Earth is cooling and the weather is cold (claims 1.2 and 1.3 in [Figure 10.1](#)), coexists with a page titled "Warming/CO2 Benefiting Earth", which asserts that the global warming and increase in carbon dioxide is beneficial for the planet, corresponding to claims 3 and 3.3. As is highlighted in [Figure 10.2](#), asserting that there is no warming (or that the Earth is cooling) and that the warming is beneficial simultaneously is contradictory, and this is but one of many examples of contradictory claims found on this blog. There are also many pages attacking the reliability of the science or climate movement (claim 5), or renewable energy technologies (claim 4).

In the pages listing hundreds of scientific references, we spotted articles from well known figures of the climate skeptics movement, such as I. M. Goklany, an electrical engineer linked with the Heartland Institute (a well-known climate skeptic think tank) or Nicola Scafetta, a physicist known for producing climate skeptics science⁴. There are also a certain number of articles which are not peer-reviewed: opinion papers, for instance, seem rather common. But, as we browsed the website, one thing became quickly apparent: it is very hard to spot whether many articles, which appear to be genuine peer-reviewed papers, actually support the skeptics claim they are meant to.

10.3.2 Misuse of articles

As a nonspecialist in climate science, it is nearly impossible to determine at first glance whether a cited article genuinely supports the claim it is being used for. The blog *NoTricksZone* sometimes provides quotes from articles that are taken out of context and filled with technical jargon. On the page titled "100K Papers - Sun Drives Climate," which aggregates about a hundred articles allegedly supporting

⁴These two examples illustrate a phenomenon which was documented by [Oreskes and Conway \(2011\)](#): the scientists promoting climate skeptics views tend to come from other domains than climate science.

the idea that climate variations are driven by the Sun, one article is cited with the sentence: “GCMs underestimated the surface LW and SW CRF and predicted near zero SW CRE when the measured values were substantially larger...” (No Tricks Zone - Sun Drives Climate, 2024). Often, however, climate skeptics simply compile lists of hundreds of articles. Even if someone had access to the full articles and the time to read them, it would still be difficult for a nonspecialist to assess their content accurately. If verifying a single paper is already challenging, the effort required to scrutinize an entire list of articles is simply out of the reach of most people.

There are documented cases of climate contrarians blatantly misusing scientific research. According to Carbon Brief (2011), several scientists whose work was featured on the list of 900 articles promoted by the Global Warming Policy Foundation disagreed with the notion that their research supported climate skepticism and had previously, without success, requested the removal of their articles from similar lists. Similarly, in an article published in the Earth science news website *Eos*, titled “What happened when climate skeptics misused my work”, a graduate student in climate science shares how his first published paper was misrepresented by climate skeptics (*Eos*, 2020). Using one sentence from the article’s abstract, contrarian Twitter accounts mistakenly claimed that the study was refuting the existence of a greenhouse effect, when it was, in fact, doing the opposite (highlighting a feedback loop between the Earth’s surface and lower atmosphere which was increasing the effect of greenhouse gas emissions). Upon reaching out to his supervisors and university administration, the graduate student learned that several of his colleagues had also experienced similar misrepresentation of their work. These examples suggest that such tactics are not uncommon among climate denial proponents.

This strategy, which can be labeled “Misuse of Science,” exploits the difficulty that understanding a climate science article poses for laypeople. As testified by the scientists cited in *Carbon Brief* and the *Eos* article, even when the authors are aware of and disagree with how their work is used to support climate skepticism, it is challenging for them to communicate this to the public targeted by climate contrarians. As a result, this approach is both cost-effective and low-risk for climate skepticism promoters: out of the thousands of readers of a blog post, many will likely be swayed by the apparent weight of scientific evidence. And for those attempting to refute these claims, verifying the accuracy of the cited research becomes a Sisyphean task. This aligns with Brandolini’s Law of misinformation (or the “bullshit asymmetry principle”), which states: “The amount of energy needed to refute bullshit is an order of magnitude larger than that needed to produce it.”

From an abstract argumentative perspective, this means that the climate contrarians can easily produce weak arguments which are time consuming to refute - a dream for opinion manipulators. This result contrasts with the way scientific debates were modeled in Chapters 6 and 7: in these models, a major hypothesis was that weak arguments were easier to attack than strong ones. In order to reflect the reality of socio-technical controversies, we may want to update this hypothesis in light of our analysis.

This qualitative analysis gave us a sense of the nature of the scientific arguments of climate contrarians. However, given the large quantity of articles referenced -we numbered dozens of web pages in the *NoTricksZone* blog alone- qualitative analysis is insufficient to fully capture the scope and composition of the referenced scientific literature. To achieve a more accurate understanding, we propose to construct a database to systematically evaluate this corpus and to more thoroughly analyze the proportions of the different types of articles referenced.

10.4 Discussion and Future Developments

In the first section of this chapter, we presented an abstract argumentation graph representing the arguments asserted by climate contrarians in their online discourse. This allowed us to identify a characteristic of climate contrarian discourse: its **self-contradiction**. This highlights the relevance of abstract argumentation as a tool to analyse discourses: like in the study presented in Chapter 9, by modelling explicitly the structure of the debate, we are able to obtain meaningful insights into the argumentative strategies.

In the second section, we reported our qualitative analysis of the scientific articles shared by a prominent contrarian blog. This analysis revealed the self-contradiction we previously identified, as well as the types of science manipulation outlined by [Oreskes and Conway \(2011\)](#) and modeled by [Weatherall et al. \(2020\)](#). Moreover, our investigation suggests that a very common strategy employed by climate contrarians is the **misuse of scientific articles**. This tactic exploits the inherent complexity of scientific articles for laypeople, misrepresenting references to support arguments that the articles themselves do not substantiate.

Thus, our observations allowed us to identify two characteristics of climate contrarian discourse: self contradiction and the misuse of scientific arguments. These characteristics reveal that climate skeptic discourse is produced by actors who, contrary to their claims, are not defending an overlooked or marginalized science. Our findings align with numerous studies that demonstrate climate skeptic discourse is largely the result of deliberate manipulation. The behavior of those promoting climate skepticism can be characterized as acting in **bad faith**, a conclusion supported the contradictions between their claims and their misuse of scientific evidence. These tactics expose a strategy aimed not at genuine debate, but at undermining the established scientific consensus on climate change.

We can note that the "misuse of science" strategy was not represented in the model of [Weatherall et al. \(2020\)](#), which primarily focused on the manipulation of legitimate scientific evidence. In their model, the articles shared by propagandists genuinely support the desired option, unlike the misuse seen in climate skeptic discourse. By misusing scientific articles, climate contrarians have a capacity to create a large number of weak arguments, which are hard to attack for laypeople. This is

reminiscent of the way biased agents can produce many more arguments than non-biased agents in the model of Chapter 7. However, neither the models of Chapter 6 nor 7 included agents who behaved in bad faith, promoting self contradictory positions. This distinction is highly informative from a modeller's perspective. These results suggest that the presence of bad faith agents in a socio-technical controversy requires tailored modeling approaches—both when constructing an appropriate argumentation framework and when developing protocols that accurately describe the dynamics introduced by such actors. This distinction underlines the need for models that account for deliberate distortions of scientific evidence, beyond the usual scope of disagreement in scientific debate.

Our study also draws attention on the importance of studying the *temporal aspect* of the debate. As demonstrated by the shifting strategies of the Global Warming Policy Foundation, and as highlighted in Coan et al. (2021), climate contrarians have changed their argumentative tactics over time. Not many people are now seriously claiming that we are "heading into an ice age", but this was a position that climate contrarians asserted very strongly a few years ago. Since contrarians change their position and stop asserting some of their arguments over time, the temporal aspect is essential to understand the position they promote and, in general, the argumentative strategies of agents debating in bad faith. This brings us back to the interest of understanding argumentation as a dynamic process, and the relevance of the use of tools such as simulation to give a dynamic account of socio-technical controversies.

In the future, we plan to enhance our analysis with empirical data collection, particularly bibliometric data. We are building a database of scientific articles cited on various climate contrarian blogs. This data will offer insights into the scale of the manipulations we identified in our qualitative analysis. For instance, inspired by the work of *Carbon Brief*, we aim to contact the authors of these referenced articles to verify whether they endorse the use of their work by climate contrarians. By doing so, we can assess whether, as we suspect, the misuse of scientific articles constitutes a significant portion of the sources cited by climate contrarians. This investigation will help to quantify the extent of bad-faith practices within the contrarian discourse and deepen our understanding of how scientific research can be manipulated.

The work presented in earlier chapters of this thesis, notably in Chapters 7 and 9, emphasized the importance of the **quantity** of arguments in debates. This focus on quantity aligns closely with what is measured by gradual argumentation semantics, such as those discussed in Chapter 5. Starting with our database, we plan to classify each scientific article according to the taxonomy of contrarian claims, as developed by Coan et al. (2021). This classification will give us a clearer picture of how many studies (according to the skeptics) support each type of claim. By collecting the number of scientific studies mobilized to back different contrarian claims, we can construct a weighted argumentation graph. In this framework, the articles would represent arguments supporting specific contrarian claims. If we gather information about the relevance or validity of these articles, we can further refine the strength

of the support relation between articles and claims.

Thus, building this database will deepen our understanding of the argumentative strategies used by climate contrarians. It will also allow us to perform a temporal analysis of the articles cited by skeptics, helping to answer key questions such as: does the science they cite evolve over time? If so, how does it change, and what does this reveal about the motivations and strategies behind climate contrarianism? By mapping the production of contrarian content to significant global events, we may gain further insights into their evolving tactics. This temporal data could also inform the development of an agent-based model.

Overall, this study reveals the complementarity of tools such as abstract argumentation, qualitative and quantitative analysis, and agent-based modelling, to improve our knowledge about socio-technical controversies in general, and our understanding of misinformation tactics in particular.

This field of study is all the more important given that misinformation strategies do not seem to be a problem of the past: on the contrary, recent research reveals that coordinated efforts to disrupt public debates are taking place around a variety of subjects. For instance, a recent investigation led by *Le Monde*, *The Guardian*, and other major media outlets has uncovered the existence of *Bonus Eventus*, a private platform used by top executives and lobbyists for agro-industries. This platform contains ready-made arguments supporting pro-pesticide positions as well as a database containing private information about hundreds of scientists and activists who could potentially influence the public debate about pesticide use. This is a clear continuation of the "tobacco strategy": years after its description by [Oreskes and Conway \(2011\)](#), industries are still using questionable tactics to delay or obstruct regulatory action. Furthermore, the misuse of scientific arguments observed in climate contrarian blogs is likely to be a current and global problem, and it may gain new tools. [Haider et al. \(2024\)](#) have recently uncovered that a significant number of scientific articles on controversial issues, such as health or the environment, were created from scratch by generative AI models like ChatGPT and listed on platforms such as Google Scholar. This suggests coordinated misinformation campaigns to produce illegitimate science. The authors of the article denote this phenomenon as "evidence hacking," which they define as "the strategic and coordinated malicious manipulation of society's evidence base." As demonstrated in this work, abstract argumentation may provide valuable insights to improve our understanding of these kinds of disinformation, a crucial task in this context of broad misuse of scientific discourse.

CHAPTER 11

Conclusion

The work presented in this thesis consists of several approaches to study socio-technical controversies with abstract argumentation, gradually moving from theoretical studies to applied work.

After an introduction to the relevant literature in Part I, Part II focused on agent-based models to explore scientific controversies. This section begins with the presentation of a novel gradual semantics to model agents' opinions in Chapter 5, followed by two agent-based models of scientific communities. In Chapter 6, agent opinions are modeled by the gradual semantics, allowing for a detailed analysis of opinion dynamics, and in Chapter 7, we focused on measuring the impact of a cognitive bias on the epistemic performance of scientists.

Part III presented two case studies of real-life socio-technical controversies: one surrounding local opposition to an industrial sawmill in Chapter 9, and the other addressing the climate change debate in Chapter 10. These empirical studies underscore the usefulness of abstract argumentation to describe argumentative strategies and giving a deeper understanding of the argumentative dynamics which characterize these controversies.

The results obtained by these different studies resonate with one another other; in particular, there are connections between the phenomena identified through multi-agent simulation and empirical studies. In this conclusion, I will explore the major findings that emerged from this body of work, addressing the initial research question regarding the impact of social structures on the circulation of arguments within socio-technical controversies. Then, I will reflect on the formal tools which were used in these studies: abstract argumentation and agent-based models. Finally, I will conclude by sketching some of the future research directions that stem from this thesis.

Results

The importance of social structures

Several works presented in this thesis highlight the importance of **social structures** in controversies: by influencing the circulation of arguments, social structures influence each individual's perception of the debate and heavily impact the outcome of the controversy. The main social structures which were explored in the agent-based

models presented in Part II of this thesis are: social networks, teams of debaters and filters of arguments. Chapter 6 demonstrated that the shape of the social network which linked scientists had a major influence on their opinions about an issue. In particular, more social connection between agents can either increase or decrease the divergence of their opinions, depending on specific conditions, a result which stands in contrast to previous ABMs of scientific communities. In Chapter 7, we found out that if we take myside bias into account, the number of people in each side at the beginning of a controversy is determinant for its outcome. Furthermore, this model demonstrated that filters of weak arguments at the level of the community could modify the outcome of a debate by lowering the number of weak arguments which enter public discussion, thus mitigating the impact of myside bias.

This importance of social structures also manifests itself in Chapter 10, which analyzes the arguments used by climate contrarians. Indeed, the social networks and blogs used by climate contrarians are crucial components of their strategy to promote climate skepticism, as was illustrated by the work of Chavalarias et al. (2023) on the network of Twitter contrarian accounts. By presenting a skewed set of arguments, contrarian actors can profoundly influence people's opinion about anthropogenic climate change. The problems which climate scientists face when they try to reach the public targeted by this misinformation further underscore the critical role of communication channels in shaping these debates.

These observations suggests that understanding controversies entails much more than merely grasping the arguments of both sides: it is necessary to know not only which arguments exists, but also **who has access to which arguments**. Therefore, an analysis of the social structures that influence how people encounter and engage with arguments is necessary. This insight reinforces the need to approach controversies as both *dynamic* and *social* phenomena. As was demonstrated in this work, both agent-based models and empirical studies are promising and complementary tools to explore these dimensions of socio-technical controversies.

Gish Gallop, bad faith and cognitive biases

In 1994, anthropologist Eugenie Scott identified an argumentative strategy called the "Gish Gallop" (Scott, 1994). Named after a prominent American creationist, Duane Gish, it consists in drowning one's opponent in a deluge of arguments which may be invalid, unverifiable, inaccurate, deliberately misleading or even irrelevant. Also known as the "firehose of falsehood", this technique rests on the fact that each argument will take much longer to refute than to assert, a principle also known as Brandolini's law (or the "bullshit asymmetry principle", which we described in Chapter 10). Thus, an opponent will not have the time to counter all of the arguments produced by the Gish Gallopper, and the public, especially if they are unfamiliar with the topic and unable to assess the arguments themselves, may be strongly influenced.

In my work, I have identified several situations, both in agent-based simulations and empirical studies, that resemble instances of the Gish Gallop. Chapter 7 demonstrated that biased agents could win a debate and convince the scientific community of their preferred position, even when it was the weaker option. This outcome was only possible when the agents supporting the weaker position were in majority at the outset of the debate. The key to this dynamic lies in the bias mechanism described by [Mercier \(2017\)](#), where myside bias allows agents to generate arguments more easily in favor of their position. Although these arguments are weaker, their sheer quantity overwhelmed the proponents of the stronger theory: unable to address them all, they were eventually swayed. This mechanism mirrors the dynamics of a Gish Gallop, though with some distinctions. The traditional definition of Gish Gallop involves two opponents debating to convince an external audience. In contrast, my model depicts a debate within a community where individuals can both generate arguments and change their opinions. In this context, the number of people defending the weaker option becomes highly significant, since the argumentation burden shifts onto the minority. This situates the Gish Gallop concept in a more social and distributed setting, where the dynamics of group size and argument production heavily influence the outcome of debates.

In the study of argumentative strategies of the proponents and opponents of a local industrial sawmill presented in Chapter 9, I identified a strategy akin to the Gish Gallop. The company promoting the project put forward a multitude of arguments that were relatively weak. These were not outright falsehoods, but they included trivial arguments and promises that were likely to be broken in the future, such as assurances regarding pollution levels (which were unlikely based on the company's track record at other locations). Rather than a "firehose of falsehood", this tactic might be better described as a "firehose of weak arguments". By overwhelming the opposition with numerous unconvincing but plausible claims, the company strengthened the impression given by their body of evidence to local public, while making it difficult for the opposing side to counter each argument effectively.

The clearest example of the Gish Gallop strategy, however, is found in the study of climate contrarian arguments presented in Chapter 10. Climate contrarians effectively mobilize scientific references to produce a large number of arguments, many of which are self-contradictory or demonstrably false. This mirrors the Gish Gallop tactic: by bombarding their audience with a volume of flawed or misleading arguments, they create doubt and make it difficult for opponents to refute each point individually. As in the model described in Chapter 7, this scenario involves more than two debaters: there is a large number of contrarian blogs and websites performing this tactic simultaneously (and, as shown by [Chavalarias et al. \(2023\)](#), sometimes in a coordinated way). The public audience consists of laypeople for whom it is nearly impossible to accurately assess the strength or validity of the scientific arguments being presented, thus making the Gish Gallop particularly effective.

These different situations may require different modeling approaches. In Chapter 9, I proposed to use **argument weights** to account for the fact that some arguments are clearly weaker than others. Using a gradual semantics may be relevant to model debates where weaker arguments can still accumulate and gain traction if they are numerous enough. In contrast, in Chapter 10, where we identified the *misuse of scientific arguments*, I suggest modeling **weights for attack relations**. This reflects that the issue is not merely the quality of the arguments themselves, but rather the **relevance of the attacks**. In these cases, climate contrarians often misinterpret or misrepresent scientific evidence. Rather than describing the contrarians' arguments as inherently weak, it is more appropriate to question the validity of the support relation they assert between the scientific evidence they provide—which may be perfectly sound—and the contrarian claims they make. Thus, these case studies support the importance of modeling both the strength of arguments and the strength of attack relations, capturing the full complexity of these socio-technical controversies.

Another important distinction lies in the *motivation* behind these strategies. In Chapter 10, the self-contradiction and deliberate misuse of scientific evidence that we observed points to bad faith. Climate contrarians knowingly use tactics like cherry-picking or taking research out of context to sow doubt. Their strategy involves intentional manipulation, and this should be reflected in a model as deliberate misinformation tactics. On the other hand, in Chapter 7, the agents are *unaware* of their bias. They may believe in the validity of their arguments, even if they're influenced by cognitive biases that make them favor their own position. Their behavior does not represent deceit, but rather cognitive features that lead to the production of biased arguments. The situation in the sawmill case (Chapter 9) is more ambiguous. It's unclear whether the executives who produced and presented the company's arguments are biased or acting in bad faith. While they could believe their own claims, it's equally possible that they are deliberately producing arguments that they know to be weak in order to protect their interests. However, regardless of their awareness, the result is the same—the opposing side is overwhelmed by a multitude of weak arguments.

The existence of these mechanisms of Gish Gallop also highlights the crucial factor of *time* in debates. Indeed, time is a necessary component of the Gish Gallop strategy: overwhelming one's opponent with a rapid succession of weak or misleading arguments works *precisely* because the opponent does not have enough time to refute each one. This creates an imbalance in the debate, allowing the weaker side to appear stronger simply by virtue of the quantity of arguments produced. This is reflected in the agent-based model of Chapter 7, where the minority agents do not have enough time to refute the arguments produced by the numerous majority agents, who can sway the debate despite the inferior quality of their claims. Similarly to our observation about social structures, we can see that controversies are much more than a simple map of arguments: rather, they evolve over time. Therefore, to fully understand and model these controversies, it is essential to represent them as dynamic processes that develop and change as the debate unfolds. This temporal

dimension is necessary to account for strategies like Gish Gallop, where the speed of argument production plays a key role.

The model presented in Chapter 7 suggests a potential method to counter the effectiveness of Gish Gallop-type strategies. In the model, a community filter that prevents weak arguments from being shared publicly reduces the advantage of a biased majority. This filter could represent mechanisms like peer-review in the scientific community, where low-quality studies are less likely to be published. However, implementing such a mechanism in broader socio-technical controversies, which involve multiple types of actors, is far more challenging. **Fact-checking** is a potential analog to this filter in the context of public debates. However, it has limitations: there is no guarantee of the independence of fact-checkers, with possible case of bias and censorship. Moreover, the analysis of Gish Gallop strategies in real-life controversies reveals a critical simplification in our model: in the simulation, the number of arguments the filter addresses is inconsequential. The filter works the same, regardless of the quantity of arguments produced. However, in reality, if a scientific community or fact-checkers are overwhelmed with a flood of weak arguments, the effort required to peer-review, refute, or fact-check them is substantial. A more realistic model would account for the constraints involved in filtering large quantities of information, potentially incorporating elements like finite resources and time available to fact-checkers or peer reviewers.

How to model socio-technical controversies

In summary, the studies of controversies presented in this manuscript have yielded the following insights about how to model socio-technical controversies:

- **Using social structures:** The way arguments circulate and shape people's opinion is heavily influenced by social structures, such as networks of actors and their communication channels. Models should represent these structures and their role in shaping the dynamics of controversies.
- **Importance of the temporal dimension:** Debates evolve over time, with strategies and arguments shifting as the controversy progresses. This necessitates dynamic modeling approaches like agent-based models to capture the way arguments are produced, communicated and abandoned throughout time.
- **Including biases and deliberate misinformation:** Socio-technical controversies are influenced not only by the cognitive biases of the participants, but also by deliberate misinformation. This dual presence should be reflected in models to account for both unintentional and strategic distortions of information.
- **Weighting arguments and attacks:** Not all arguments are of equal worth, and the relevance or strength of attack relations can vary. Models should

include weighted arguments and attacks if they aim to reflect biases and bad faith argumentative strategies.

- **Using gradual semantics:** Finally, the studies suggest that models should employ gradual semantics to capture the cumulative impact of arguments, which is central in the argumentative strategies that we identified.

Insights about the tools

This section summarizes reflections on what this research can teach us about the formal tools that contributed to it.

Abstract argumentation

My work illustrates the usefulness and flexibility of abstract argumentation. Firstly, abstract argumentation has proved very valuable for modeling agent cognition in a simple yet informative way. Secondly, when used in the context of real-life case studies, it has proven to be a fruitful means of representing the structure of controversies. Using it to describe real-life controversies allows us to identify and understand distinct argumentative characteristics and strategies. In particular, thanks to formal tools concerning the acceptability of arguments, we were able to formally identify the self-contradiction of climate opponents in the Chapter 10. This enabled us to substantiate our claim that this position is not rational, underlining the usefulness of formal approaches in case study analysis.

When I presented abstract argumentation frameworks to participants in the Chapter 9, I remarked that this enabled a certain reflexivity and improved understanding of the two positions in the debate. This can help improve the argumentative strategies of the actors involved. In other cases, it can help develop an understanding of the opposing side's position, which can be useful in reaching a consensus.

As developed above, my work supports the use of weighted and gradual approaches to abstract argumentation, in line with many other empirical works. In conclusion, abstract argumentation is a promising tool in the study of socio-technical controversies. In particular, it is fruitful when combined with approaches such as agent-based models, or when complemented by data collection.

Agent-based models

Agent-based models enable us to study the dynamic and social aspects of debates which, as we have shown, are essential components of socio-technical controversies. The modeling work presented in Part II has contributed to understanding the complexity of socio-technical controversies. These models have highlighted the importance of social structures, and the modeling of myside biases in particular has enabled a better understanding of the phenomena identified in the case studies. As developed above, my work identifies the presence of deliberate misinformation in controversies. As such, it recommends modeling heterogeneous agents whose motivations may be distinct.

Future directions of research

Completing the work on case studies

The studies of real-life cases of controversies presented in Chapter 9 and Chapter 10 are works-in-progress. The first direction of my research will be to complete these works, enhancing the argumentation frameworks representing each controversy. The work presented in Chapter 9 will be completed by adding weights to the arguments and applying a gradual semantics to assess each argument's acceptability. This will give further insights into the strengths and weaknesses of each position. In Chapter 10, we plan to complete the database of scientific articles mobilized by climate skeptics, gathering further insights about the number of scientific arguments which support each of the contrarian's claims, to construct a detailed framework of the controversy.

One thing to note about these two controversies is that they are of very different scales: the debate studied in Chapter 9 mainly involves local actors and is rather recent, whereas the climate change controversy has been going on at an international level for many years, involving decades of scientific knowledge and hundreds of thousands of actors. A more in-depth study of comprehensive argumentation frameworks could enable us to understand how the scale of a controversy influences its structure, dynamics and the argumentation strategies deployed. In the future, to complement these analyses, I would like to study other controversies, perhaps of a different scale or theme, to explore differences and similarities in structures, strategies or arguments.

Developing an agent-based model of socio-technical controversies

Another, rather natural, future direction for this research would be to exploit the insights gained from my empirical studies to create an agent-based model representing broader socio-technical controversies. Inspired by the scientific debate models that I have already created, such a model would implement the features described above. It would involve weightings for arguments and attacks, and heterogeneous agents, taking into account the presence of attempts at manipulation with (coordinated) agents debating in bad faith. This would allow us to explore the dynamics induced by deliberate misinformation strategies. As we did with the bias filter in Chapter 7, we could also test ways of mitigating them. We could study the interaction between cognitive biases and strategic misinformation. For example, we could investigate the existence of feedback loops in which cognitive biases and social phenomena such as homophily reinforce the effectiveness of misinformation.

Exploring the links between my work and Actor-Network-Theory

Another promising area of research will be to further explore the meaningful connections between abstract argumentation and Actor-Network Theory (ANT), like we began to do with the tool developed in Chapter 9. ANT is a well-studied and fruitful approach for analyzing socio-technical controversies. As a computer scientist, my background has not yet allowed me to fully engage with this body of work, but comparing both approaches and results would be of the utmost interest for assessing the power of abstract argumentation to study controversies. We could verify whether my conclusions match with those obtained with ANT. For instance, in Chapter 9, I noted that participants who could reflect on their own argumentative strategies tended to modify their positions. This is reminiscent of one of the major results of ANT, which is that controversies, rather than revealing existing oppositions, are situations that lead to the creation and developments of opposing sides throughout the conflict (Akrich et al., 2006). Furthermore, one of the key features of ANT is that it treats both humans and non-humans -who can be concepts, technologies, institutions, or material objects- as actors that influence the network of relations within a controversy. On the contrary, opinion dynamic models in general, and those using abstract argumentation in particular, typically focus on human agents and their interactions. Thus, drawing from ANT methodology, we could enhance our models by incorporating non-human actors when modeling socio-technical controversies. This integration would raise interesting questions, such as how non-human actors can be represented in abstract argumentation frameworks, and how their influence on the circulation and strength of arguments can be quantified. Additionally, it could lead to the development of new types of models that better capture the complexity and interconnectedness of real-world controversies, extending the applicability of agent-based models and argumentation theory in the study of socio-technical dynamics.

Popularizing and assessing the impact of my research

After completing my PhD, I will be working on popularizing some of my results to the public. Although this is not exactly part of academic research, popularizing results can be extremely instructive and enhance the usefulness of research for civil society. In particular, the study of the controversy surrounding the proposed mega-sawmill at Saint-Agnan involved a link with the local public, which makes its communication highly relevant. Recently, local environmental activists, interested in the sawmill project but unfamiliar with it, contacted me to ask for a summary of the arguments on both sides of the controversy. I realized that my work did not yet offer a simple way of visualizing and understanding the positions of both sides, especially for lay people unfamiliar with programming or abstract argumentation. Yet the interest in this research and the impact it could have on local people's knowledge is obvious. That's why I'd like to explore ways of communicating controversies to the public, possibly using abstract argumentation. For example, this could take the form of argumentation maps summarizing the controversy in a simple, visual way. This would also increase our knowledge of the usefulness of abstract argumentation in communication. This brings to mind the work of [Vesic et al. \(2022\)](#), who studied the effect of visual representation of argumentation graphs on people's cognition.

This focus on popularization is part of a wider context of personal reflections on the use of science and the role it plays in the profound transformations facing our world. On a personal level, I believe that the global environmental crisis we are facing, which threatens the conditions that make our planet a liveable environment for us and all its other inhabitants, should be at the heart of our efforts as scientists. These questions have led me to participate in several collective efforts to reflect on the role that science (and in particular computer science) plays, and could play, in the situation we face. This attention has been reflected in my choice of subjects, and I hope to continue along this path, exploring how science in general, and my work in particular, can relate to efforts on the ground to understand and fight the causes of this crisis.

In summary, the work presented in this manuscript can be seen as the beginning of a broader research project, using abstract argumentation in different ways to explore and understand socio-technical controversies. Given the variety of issues that abstract argumentation has already enabled us to address, I am confident that further exploration of the research questions outlined by this work will bring interesting new insights and results. What's more, this work will involve a fascinating interplay between various fields of study, such as formal argumentation, agent-based models, bibliometrics, or actor-network theory. At the same time, it will contribute to our understanding of the pressing current issues raised by scientific misinformation.

Remerciements

Tout d'abord, un très grand merci à mes encadrantes de thèse, Juliette Rouchier et Gabriella Pigozzi. Il m'est difficile d'imaginer un meilleur encadrement de thèse. Non seulement pour vos conseils avisés, vos réflexions pertinentes, vos mots justes et votre fine connaissance du monde académique, mais aussi pour votre patience, votre écoute et votre gentillesse, tellement nécessaire pour m'accompagner dans cette découverte de la recherche parfois mouventée. Merci d'avoir considéré mes idées avec ouverture et bienveillance et d'avoir accompagné mon évolution au cours de ces trois années. Je me sens très chanceuse d'avoir pu travailler avec vous.

Ensuite, j'aimerais remercier Nicolas Maudet et Elise Bonzon, qui ont encadré mon premier travail sur l'argumentation, qui m'ont mis sur la voie de cette thèse et qui ont aujourd'hui accepté d'en être jury. Merci également à Srdjan Vésic et à Pascale Zaraté, pour nos discussions et votre travail de rapporteur, qui m'a aidé à réaliser ce manuscrit, et à Sylvie Huet, Rallou Thomopoulos et Paola Cantù d'être jury de ma thèse.

Au sein du monde académique, j'ai eu la chance de profiter des conseils éclairés de Florence Dupin de Saint-Cyr, qui m'a invitée à l'IRIT à Toulouse et d'Antony Hunter qui m'a accueillie à l'University College London. Ces séjours ont affiné mes réflexions et ma recherche. Merci à mes co-auteurs; Matteo Michellini, AnneMarie Borg, Dunja Seselja et Christian Strasser, de m'avoir inspiré par votre rigueur et votre exigence. Je voudrai remercier tout particulièrement Matteo, qui m'a parlé le premier du *Gish Gallop*. J'ai profité de discussions éclairantes avec d'autres doctorants travaillant sur l'argumentation, en particulier Jordan Thieyre. Et merci aux doctorants du Lamsade avec qui j'ai partagé de nombreux moments joyeux et studieux.

Je dois aussi beaucoup à tous.tes les chercheurs et chercheuses avec qui j'ai pu parler du lien entre recherche et écologie politique : notamment Matthieu Latapy, Pacôme Delga, et les autres membres du groupe de réflexion de l'Ecopolien.

Bien sûr, cette thèse n'aurait pu avoir lieu dans mes ami.es et ma famille, de précieux soutiens à toutes les étapes de cette aventure.

Merci à et Adam, avec qui j'ai parlé en premier de faire une thèse. Merci à Louis, pour ton soutien inconditionnel pendant une période cruciale. Merci à Juliette et François pour toutes les découvertes, les discussions et l'accueil. Un merci tout particulier à Tiphaine avec qui j'ai vécu en parallèle la thèse et la découverte d'une vie associative riche et passionnante: ton enthousiasme et ton énergie ont changé la trajectoire de ces trois dernières années. Merci à Maxime, pour m'avoir intégrée au projet, pour nos discussions sur la science, et pour être dans ma vie. Merci à mes colocataires, les anciens des Gens qui Sèvent, et les nouveaux des Radis, pour m'avoir écoutée, soutenue, encouragée et fait à manger pendant les moments les

plus difficiles.

Merci à mes grand-parents : en particulier à ma grand-mère Marie chez qui j'ai pu trouver un havre pour travailler et de reposer, et à ma grand-mère Josiane qui m'a toujours accompagnée dans mes étapes de vie et qui n'a malheureusement pas pu aller jusqu'au bout de celle ci.

Merci à mes parents qui ont beaucoup fait pour que je puisse poursuivre des études exigeantes, et à mes frères et soeurs Antoinette, Ferdinand et Josse, toujours de fidèles soutiens et repères de ma vie.

Acknowledgements

First of all, a very big thank you to my thesis supervisors, Juliette Rouchier and Gabriella Pigozzi. It's hard to imagine better thesis supervisors. Not only for your sound advice, your pertinent reflections, your apt words and your fine knowledge of the academic world, but also for your patience, your attentiveness and your kindness, so necessary to accompany me in this sometimes turbulent discovery of research. Thank you for considering my ideas with openness and kindness, and for accompanying my evolution over these three years. I feel very lucky to have been able to work with you.

Next, I'd like to thank Nicolas Maudet and Elise Bonzon, who supervised my first work on argumentation, who set me on the path to this thesis and who have now agreed to be on its jury. Thanks also to Srdjan Vésic and Pascale Zaraté, for our discussions and your work as rapporteurs, which helped me to complete this manuscript, and to Sylvie Huet, Rallou Thomopoulos and Paola Cantù for being in my PhD jury.

In the academic world, I was fortunate to benefit from the enlightened advice of Florence Dupin de Saint-Cyr, who invited me to IRIT in Toulouse, and Antony Hunter, who hosted me at University College London. These stays have sharpened my thinking and my research. Thank you to my co-authors; Matteo Michellini, Anne-Marie Borg, Dunja Seselja and Christian Strasser, for inspiring me with your rigor and high standards. I'd especially like to thank Matteo, who first told me about *Gish Gallop*. I benefited from enlightening discussions with other PhD students working on argumentation, in particular Jordan Thieyre. And many thanks to the Lamsade PhD students with whom I shared happy and studious moments.

I also owe a debt of gratitude to all the researchers with whom I was able to discuss the link between research and political ecology: in particular Matthieu Latapy, Pacôme Delga, and the other members of the Ecopolien.

Of course, this thesis could not have taken place without my friends and family, who have been invaluable supporters at every stage of this adventure.

Thanks to Adam with whom I first talked about doing a PhD. Thanks to Louis, for your unconditional support during a crucial period. Thanks to Juliette and François for all the discoveries, discussions and hospitality. A special thank you to Tiphaine, with whom I experienced both the thesis and the discovery of a rich and exciting activist life: your enthusiasm and energy have changed the trajectory of these last three years. Thanks to Maxime, for integrating me into the project, for our discussions on science, and for being part of my life. Thank you to my roommates, the old ones from Les Gens qui Sèvent, and the new ones from Les Radis. Thank you for listening to me, supporting me, encouraging me and providing me with food during the most difficult times.

Thank you to my grandparents: in particular to my grandmother Marie, with whom I was able to find a haven to work and rest, and to my grandmother Josiane, who had always accompanied me through all the stages of my life and who unfortunately has not been able to go to the end of this one.

I would also like to thank my parents, who did much to enable me to pursue my demanding studies, and my brothers and sisters Antoinette, Ferdinand and Josse, who have always been faithful supports and landmarks in my life.

Résumé de la Thèse en Français

Introduction et contexte

Voudriez-vous boire un bol de RoundUp ? Étonnamment, la science n'a pas tranché la question de la toxicité du glyphosate, principal composant actif du célèbre herbicide de Monsanto. Cette controverse a atteint son paroxysme en 2015 avec la classification de la molécule comme « probablement cancérigène » par le Centre international de recherche sur le cancer, avec des preuves contradictoires pointant vers les deux camps, et la révélation des « Monsanto papers », une collection de documents internes montrant la corruption des scientifiques par l'entreprise. Elle a notamment menacé la réautorisation du pesticide par l'UE en 2022, une décision qui aurait pu entraîner un changement majeur des pratiques agricoles en Europe ([Kudsk and Mathiassen, 2020](#)).

Les débats scientifiques comme ceux sur le glyphosate ont attiré l'attention des chercheurs depuis les années 1980, à la suite des travaux des anthropologues des sciences Bruno Latour et Steve Woolgar, ce qui a conduit à l'émergence du domaine des **controversy studies** (études sur les controverses). ([Latour and Woolgar, 1979b](#)). [Lemieux \(2007\)](#) propose une définition de la controverse comme un débat caractérisé par une structure triadique, où deux camps opposés tentent de convaincre un « public » réel ou virtuel, et qui se déroule dans un environnement plutôt fermé de spécialistes.

Cependant, comme nous pouvons le voir dans le cas du glyphosate, certains débats scientifiques vont au-delà d'une simple dispute entre spécialistes et impliquent une grande variété d'acteurs, des activistes aux industriels, en passant par les journalistes ou les décideurs politiques. Dans la littérature, les **controverses sociotechniques** sont apparues comme un sujet d'intérêt, avec une signification plus large : ce terme englobe les controverses qui traversent la société et s'appuient sur des arguments de nature scientifique ou technique. Elles mobilisent typiquement des notions scientifiques, techniques, juridiques, morales et politiques.

Illustrons notre propos par un autre exemple, très étudié, de controverse socio-technique : le débat sur les voitures électriques. Ce débat oppose un camp "pro" composé en partie de gouvernements et d'acteurs pro-technologiques affirmant que les voitures électriques sont moins polluantes et plus écologiques que les voitures à essence, et un camp "contre", composé de certains groupes écologistes mais aussi de théoriciens du complot ([Wagener, 2024](#)). Les écologistes contestent notamment l'aspect écologique de ces voitures en raison du coût environnemental de la fabrication des batteries. Bien qu'elle n'ait guère débuté comme un débat entre scientifiques, cette controverse mobilise des arguments scientifiques, par exemple concernant les taux d'émission de CO₂, et des arguments techniques, comme ceux concernant les performances de la voiture, mais également des arguments moraux,

critiquant les conditions de travail inhumaines entourant l'exploitation des minerais pour les batteries électriques.

La théorie des réseaux d'acteurs, développée par Bruno Latour, Michel Callon et Margaret Akrich dans les années 1980, est une méthodologie d'étude des controverses socio-techniques (Akrich et al., 2006). Cette méthode consiste à identifier les acteurs d'une controverse et les réseaux qui les relient : l'originalité de l'ANT est que les humains et les non-humains, tels que les concepts, les objets ou les animaux, peuvent être considérés comme des acteurs du réseau. A l'instar des travaux de Latour and Woolgar (1979a), elle met l'accent sur l'impact des caractéristiques sociales sur l'issue des controverses. Les controverses socio-techniques étant devenues un sujet d'étude à part entière, d'autres méthodes sont utilisées pour les étudier. Par exemple, (Wagener, 2024) s'appuie sur l'analyse du discours et les sciences humaines numériques pour caractériser la controverse sur les voitures électriques.

Au début des années 2000, le sociologue François Chateauraynaud a développé un programme informatique appelé Prospéro (PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur) pour explorer les controverses publiques (Chateauraynaud et al., 2003). Cet outil permet d'analyser les données textuelles relatives aux controverses afin de caractériser les formes d'expression et d'argumentation utilisées par les acteurs, et se veut une aide pour les sociologues qui étudient des controverses caractérisées par une grande quantité de textes. Il s'agit d'une des premières approches utilisant des outils informatiques pour explorer les controverses socio-techniques, ici en conjonction avec des méthodes sociologiques, et qui a donné naissance au champ d'étude de la "sociologie argumentative des controverses".

Parallèlement, dans le domaine de l'informatique, les dernières années ont vu le développement de l'argumentation formelle comme méthode d'étude des débats. Dans mon travail, je propose d'étudier les controverses socio-techniques à l'aide de l'argumentation formelle. Je m'intéresse particulièrement à la **circulation des arguments** ; en particulier, j'étudie le rôle des **structures sociales** sur la circulation des arguments au sein de ces controverses. En étudiant ces aspects d'un point de vue informatique, j'utilise **simulation sociale** pour explorer la dynamique sociale des controverses. En tant que telle, ma thèse s'inscrit dans plusieurs littératures distinctes, que je vais maintenant présenter avant d'aborder plus en détail ma question de recherche et mes méthodes.

Argumentation formelle

L'argumentation formelle se situe à l'intersection de la logique et de l'informatique, impliquant des méthodes de représentation, d'analyse et d'évaluation des arguments à l'aide de structures mathématiques formelles. Cette approche permet de modéliser précisément les processus argumentatifs et de développer des algorithmes capables

de simuler, d'évaluer, voire d'automatiser les tâches de raisonnement. En traitant les arguments comme des objets formels, l'argumentation formelle facilite la modélisation et la compréhension des débats.

L'un des courants les plus influents de l'argumentation formelle est l'argumentation abstraite, introduite par Phan Minh Dung dans son article fondateur sur l'acceptabilité des arguments (Dung, 1995). L'argumentation abstraite se concentre sur les relations entre les arguments, en particulier sur la manière dont ils se soutiennent ou s'attaquent mutuellement, plutôt que sur la structure interne ou le contenu des arguments eux-mêmes. Le cadre de Dung simplifie la nature complexe des interactions argumentatives en les représentant comme des objets abstraits. Ainsi, les débats peuvent être représentés comme des graphes, où les nœuds représentent les arguments et les arêtes les relations d'attaque entre eux. Dans la proposition de Dung, il existe un seul type de relation entre les arguments : la relation d'attaque.

Prenons par exemple l'échange d'arguments suivant : imaginons que deux personnes, Greta et Carole, discutent de la construction d'une nouvelle autoroute. Carole commence par dire : « La nouvelle autoroute est nécessaire car elle reliera un territoire rural plutôt pauvre au reste de la région ». Greta répond : « Il y a déjà une route et une ligne de chemin de fer, on pourrait les moderniser pour qu'elles servent à cet usage. » Cette discussion peut être comprise comme étant composée de deux arguments distincts : a : « La nouvelle autoroute est nécessaire car elle reliera un territoire rural plutôt pauvre au reste de la région » et b : « Il y a déjà une route et une ligne de train, elles pourraient être modernisées pour servir cet objectif », où l'argument b attaque l'argument a . Cette discussion peut être représentée dans un graphe d'argumentation tel que celui présenté dans la figure 11.1.



Figure 11.1: Graphe d'argumentation représentant les deux arguments échangés par Carole et Greta.

This abstraction allows for a powerful and flexible way to analyze argumentative dynamics. As such, abstract argumentation seems a promising tool to explore socio-technical controversies. However, until now, few works have attempted to study socio-technical controversies from an formal argumentative point of view.

Šešelja and Straßer (2013) appliquent l'argumentation abstraite pour représenter les débats scientifiques, proposant d'améliorer les cadres d'argumentation abstraite en ajoutant une relation « explicative ». Ce travail fournit des informations précieuses sur la structure des débats scientifiques : par exemple, ils montrent que les théories scientifiques peuvent être modélisées par des ensembles d'arguments. Nous notons que les auteurs concentrent leur analyse exclusivement sur les controverses ayant lieu entre scientifiques, sans l'étendre à l'ensemble plus large d'acteurs qui caractérisent les controverses socio-techniques. Salliou and Thomopoulos (2018) utilisent un cadre d'argumentation abstrait pour représenter le débat autour des

régimes végétariens. Ils identifient les arguments réels qui forment le débat et construisent un graphe d'argumentation explicite montrant les relations d'attaque entre eux. Cette représentation graphique leur permet d'identifier les arguments clés de la controverse, qui occupent des positions centrales dans le graphe.

Ces études suggèrent que la formalisation des controverses à l'aide d'une argumentation abstraite est un moyen d'explorer la *structure* des débats, en fournissant des informations qui peuvent compléter les études non formelles. Les controverses sociotechniques sont toutefois caractérisées par des aspects dynamiques et sociaux, qui ne peuvent être représentés par la seule argumentation abstraite. Pour inclure ces aspects, l'argumentation abstraite peut être complétée par des **modèles multi-agents**.

Modèles multi-agents

Les modèles multi-agents (ABM) sont une classe de modèles qui représentent des agents individuels, spécifient les interactions entre eux et étudient le résultat de ces interactions. Les modèles à base d'agents mettent en œuvre une approche ascendante : plutôt que de supposer une structure globale prédéterminée, les règles sont conçues au niveau des agents individuels. Ils sont particulièrement bien adaptés à l'étude de phénomènes complexes et non linéaires tels que les interactions sociales. Une caractéristique importante des ABM est leur capacité à modéliser l'hétérogénéité : les agents des ABM peuvent avoir des caractéristiques, des règles et des comportements différents. Cela permet de modéliser diverses populations ou entités au sein d'un système.

Les modèles basés sur des agents sont issus de l'informatique, mais leur histoire est riche et s'étend à plusieurs disciplines : il s'agit d'un outil multidisciplinaire, utilisé notamment en économie, en biologie, en philosophie et en sociologie. En 1971, Thomas Schelling a développé l'un des premiers modèles basés sur des agents, pour étudier la ségrégation raciale ([Schelling, 1971](#)). Son modèle utilisait des agents individuels prenant des décisions basées sur des préférences locales, et a permis de montrer comment des préférences même légères pour des voisins de la même couleur pouvaient conduire à une ségrégation significative. Ce modèle a permis de mettre en évidence le concept clé de **comportement émergent** dans les systèmes sociaux : un modèle présente un comportement émergent lorsque la dynamique collective du système n'est pas explicitement programmée mais résulte des interactions locales des agents. Ils sont bien adaptés à l'étude des situations de communication entre humains, qui consistent en de nombreuses interactions locales, et sont donc utilisés dans le domaine de la **dynamique des opinions** pour étudier l'évolution des opinions des agents.

Récemment, des modèles basés sur des agents ont été appliqués pour modéliser

des débats, en conjonction avec les théories de l'argumentation formelle. En particulier, des modèles tels que (Gabbriellini and Torroni, 2014; Butler et al., 2019; Dupuis de Tarlé et al., 2022a) utilisent l'argumentation abstraite pour représenter la cognition des agents, qui peuvent former des opinions et échanger des informations en utilisant des arguments abstraits. Cela permet de construire des modèles de diffusion d'opinion améliorés pour étudier les dynamiques argumentatives. Par exemple, Borg et al. (2017a) représente un débat entre scientifiques, et Taillandier et al. (2021) modélise la diffusion de régimes végétariens dans une population par la diffusion d'arguments en faveur ou contre la consommation de viande.

À ma connaissance, aucun modèle basé sur des agents ne modélise explicitement les controverses sociotechniques. Les modèles basés sur des agents qui étudient les controverses sociotechniques font partie d'une littérature connexe : **l'épistémologie sociale computationnelle**.

Epistémologie sociale

La philosophie computationnelle est l'utilisation d'ordinateurs (en particulier de simulations) pour éclairer les questions philosophiques. L'épistémologie est l'étude des croyances et des connaissances des individus et, en particulier, de ce qui constitue une bonne connaissance. L'épistémologie sociale étudie la manière dont les normes, les structures et les interactions au niveau du groupe influencent les croyances et les connaissances des individus.

Historiquement, l'épistémologie a été étudiée par de nombreux philosophes tels que Hume, Descartes ou Kant d'un point de vue *individuel*. L'épistémologie classique se concentre sur un individu isolé et sur ses moyens de former des croyances et d'acquérir des connaissances. En effet, la plupart de ces travaux sont issus du siècle des Lumières, une époque philosophique qui mettait l'accent sur les individus et leurs capacités rationnelles. Au cours des 20 dernières années, une branche de l'épistémologie s'est recentrée sur les aspects *sociaux* de l'épistémologie. L'épistémologie sociale reconnaît que les individus font fondamentalement partie de groupes sociaux et qu'ils acquièrent leurs connaissances par le biais d'interactions déterminées par les structures sociales. L'épistémologie sociale étudie la manière dont les normes, les structures et les interactions au niveau du groupe influencent la formation des opinions, des croyances et des connaissances. Alvin I. Goldman, un précurseur dans ce domaine, a soutenu dans son livre *Knowledge in a Social World* que les modèles basés sur des agents étaient des outils appropriés pour étudier ces questions (Goldman, 1999).

Sur la base de ses contributions, des modèles de communautés épistémiques basés sur des agents ont été développés pour explorer ces questions philosophiques. Ces modèles sont souvent similaires aux modèles de dynamique d'opinion, mais comportent une caractéristique supplémentaire : les agents peuvent formuler des opinions

qui sont épistémologiquement meilleures - c'est-à-dire plus proches de la vérité - que les autres. En examinant le « succès épistémique » des agents, ils peuvent répondre à des questions telles que : Quels types de réseaux sociaux conduisent à des taux de réussite plus élevés lorsque les scientifiques choisissent entre deux alternatives ? La désinformation, la propagande ou les médias biaisés faussent-ils la formation de l'opinion ? Kevin Zollman, Cailin O'Connor et leurs collègues ont exploré ces questions à l'aide de modèles bayésiens, dans lesquels les agents étudient et mettent à jour leurs préférences selon des lois de probabilité bayésiennes (Zollman, 2007; O'Connor and Weatherall, 2018). Notamment, O'Connor and Weatherall (2019) a créé un modèle basé sur des agents pour étudier les controverses sociotechniques, en incorporant non seulement les scientifiques mais aussi les décideurs politiques, les propagandistes et les médias dans le modèle. Cela leur a permis d'étudier comment la manipulation de la science façonne l'opinion publique. Ces modèles sont devenus très influents dans l'épistémologie sociale computationnelle, et ils ont inspiré mes propres modèles basés sur l'argumentation.

À ma connaissance, avant mon travail, il n'existait qu'un seul modèle basé sur des agents qui utilisait l'argumentation abstraite pour représenter une communauté scientifique : le modèle ArgABM de Borg et al. (2017a). Ce modèle se concentre exclusivement sur les débats entre scientifiques ; aucune tentative n'a été faite pour étudier les controverses sociotechniques à l'aide de l'argumentation abstraite au-delà du contexte scientifique.

*Question de recherche et méthode

L'objectif de ma thèse est d'**explorer comment l'argumentation abstraite peut nous aider à comprendre la dynamique des controverses sociotechniques**. En particulier, je m'intéresse à la manière dont **les structures sociales** influencent la **circulation des arguments** au sein de ces controverses. Dans mon travail, j'explore cette question en utilisant plusieurs approches : à la fois théoriques et appliquées, en me concentrant soit sur l'aspect scientifique, soit sur l'aspect sociopolitique des controverses.

Mon point de départ pour répondre à cette question a été l'étude d'un débat entre scientifiques. J'ai développé deux modèles de communautés scientifiques basés sur des agents, utilisant l'argumentation abstraite pour représenter leur cognition et la circulation des arguments dans les espaces sociaux et publics. Ces modèles sont innovants tout en s'inscrivant dans la tradition de l'épistémologie sociale formelle et, plus largement, de la dynamique d'opinion. En particulier, j'ai étudié le rôle des structures sociales : la forme du réseau social, et la présence de filtres d'arguments faibles, sur l'issue des débats scientifiques. Ce dernier élément a été utilisé dans un modèle que nous avons développé pour étudier l'impact des biais cognitifs dans le raisonnement scientifique. En développant ces modèles, j'ai découvert que la fonction que j'utilisais pour évaluer la force des arguments, l'*sémantique graduelle*, manquait de certaines propriétés nécessaires pour une simulation ABM : cela m'a conduit à proposer une nouvelle sémantique graduelle.

Dans la dernière partie de ma thèse, j'ai étudié des cas réels de controverses socio-techniques afin de tester le pouvoir de représentation de l'argumentation abstraite pour les controverses socio-techniques. En effet, compte tenu de l'absence de tentatives antérieures, il était essentiel de recueillir des informations empiriques sur la manière dont ces débats pouvaient être représentés à l'aide d'une argumentation abstraite.

Le premier cas d'application que j'ai étudié est un débat entre les partisans et les opposants à la construction d'une grande scierie dans le Tarn, en France. Les luttes contre la construction de GPII (« Grands Projets Inutiles et Imposés ») ont récemment fait l'objet d'une grande attention de la part des médias en France ; ce sont des cas évidents de controverses socio-techniques, bien qu'elles diffèrent des débats centrés sur la science à plusieurs égards. J'ai choisi ce cas comme premier test parce qu'il était représentatif, tout en étant plus petit et plus facile à traiter que les grands débats nationaux. J'ai construit un outil de visualisation et mené des entretiens qui m'ont permis de représenter la controverse sous la forme d'un graphique d'argumentation abstraite, en recueillant des informations sur les stratégies d'argumentation des deux camps.

Comme deuxième cas d'utilisation, j'ai étudié un exemple bien connu de controverse socio-technique : le débat sur le changement climatique. Je me suis concentrée sur l'utilisation d'arguments scientifiques par les blogs climato-sceptiques. En utilisant l'argumentation abstraite et l'analyse qualitative, j'ai montré leurs contradictions internes et l'accumulation de « preuves » scientifiques comme stratégie d'argumentation.

Ainsi, ma thèse peut être lue comme un passage progressif des études théoriques vers les applications : en commençant par des idées sur les propriétés des outils formels d'argumentation abstraite, suivies par un travail de simulation, et se terminant par l'étude de controverses de la vie réelle. Ensemble, ces travaux brossent un tableau des controverses comme des processus dynamiques, sujets à la manipulation et dont l'issue est fortement influencée par les structures sociales. Cela correspond à plusieurs conclusions de la sociologie des controverses, montrant la pertinence des outils formels, et en particulier de l'argumentation abstraite, pour étudier ces questions.

Structure de la thèse

Cette thèse est divisée en trois parties. La première partie, "Contexte et outils", présente le contexte nécessaire : le chapitre 2 présente l'argumentation abstraite plus en détail et le chapitre 3 les modèles basés sur les agents, y compris l'épistémologie sociale. Le chapitre 4 propose une comparaison approfondie des modèles basés sur les agents qui utilisent l'argumentation abstraite.

La deuxième partie de cette thèse, intitulée "Modéliser les controverses scientifiques avec l'argumentation abstraite", présente l'essentiel de mon travail : Le chapitre 5 introduit un cadre général pour modéliser l'opinion des agents avec l'argumentation abstraite, qui est appliqué dans le chapitre 6, où je présente un modèle à base d'agents d'une communauté scientifique. Le chapitre 7 est un autre modèle de communauté scientifique, utilisant un paradigme différent, qui étudie l'impact d'un biais cognitif sur la performance des scientifiques.

La troisième partie, "Exploration des applications", présente deux travaux en cours qui appliquent les résultats et les outils développés dans les chapitres précédents à des cas d'utilisation spécifiques. Dans le chapitre 9, je présente un outil qui utilise l'argumentation abstraite pour comprendre et informer les stratégies d'un groupe d'opposants à la construction d'un projet industriel. Le chapitre 10 se concentre sur la controverse sur le changement climatique, en présentant une caractérisation des arguments scientifiques mobilisés dans les plates-formes sceptiques sur le climat.

Le travail présenté dans ma thèse a été publié comme suit : [Dupuis de Tarlé et al. \(2023\)](#) dans les actes de la 19ème Social Simulation Conference (SCC), [Dupuis de Tarlé \(2023\)](#) dans les actes des Journées d'Intelligence Artificielle Fondamentales (JIAF), [Dupuis de Tarlé et al. \(2024\)](#) dans le Journal of Artificial Societies and Social Simulation (JASSS). [Dupuis de Tarlé et al. \(2023\)](#) a reçu le Best Student Paper Award de la Social Simulation Conference. Certaines des descriptions des modèles argumentatifs basés sur des agents présentés dans la partie I seront publiés en tant que [Dupuis de Tarlé et al. \(à paraître\)](#) dans le Handbook of Formal Argumentation Volume 3.

Résultats, discussion et conclusion

Le travail présenté dans cette thèse consiste en plusieurs approches pour étudier les controverses socio-techniques avec une argumentation abstraite, passant progressivement d'études théoriques à des travaux appliqués.

Après une introduction à la littérature pertinente dans la partie ??, la partie ?? s'est concentrée sur les modèles basés sur les agents pour explorer les controverses scientifiques. Cette section commence par la présentation d'une nouvelle sémantique graduelle pour modéliser les opinions des agents au chapitre 5, suivie de deux modèles de communautés scientifiques basés sur des agents. Dans le chapitre 6, les opinions des agents sont modélisées par la sémantique graduelle, ce qui permet une analyse détaillée de la dynamique des opinions, et dans le chapitre 7, nous nous sommes concentrés sur l'étude de l'impact d'un biais cognitif sur la performance épistémique des scientifiques.

La partie III a présenté deux études de cas de controverses socio-techniques

réelles : l'une portant sur l'opposition locale à une scierie industrielle dans le chapitre 9, et l'autre sur le débat sur le changement climatique dans le chapitre 10. Ces études empiriques soulignent l'utilité de l'argumentation abstraite pour décrire les stratégies argumentatives et pour mieux comprendre les dynamiques argumentatives qui caractérisent ces controverses.

Les résultats obtenus par ces différentes études entrent en résonance les uns avec les autres ; en particulier, il existe des liens entre les phénomènes identifiés par la simulation multi-agents et les études empiriques. Dans cette section, j'explorerai les principaux résultats qui ont émergé de ce corpus de travail, en répondant à la question de recherche initiale concernant l'impact des structures sociales sur la circulation des arguments au sein des controverses socio-techniques. Ensuite, je réfléchirai aux outils formels qui ont été utilisés dans ces études : l'argumentation abstraite et les modèles à base d'agents. Enfin, je conclurai en esquisant quelques directions de recherche futures qui découlent de cette thèse.

Résultats

L'importance des structures sociales

Plusieurs travaux présentés dans cette thèse soulignent l'importance des **structures sociales** dans les controverses : en influençant la circulation des arguments, les structures sociales influencent la perception du débat par chaque individu et ont un impact considérable sur l'issue de la controverse. Les principales structures sociales qui ont été explorées dans les modèles basés sur des agents présentés dans la partie II de cette thèse sont : les réseaux sociaux, les équipes de débatteurs et les filtres d'arguments. Le chapitre 6 a démontré que la forme du réseau social qui relie les scientifiques a une influence majeure sur leurs opinions. En particulier, une plus grande connexion sociale entre les agents peut soit augmenter soit diminuer la divergence de leurs opinions, en fonction de conditions spécifiques, un résultat qui contraste avec les précédents ABMs des communautés scientifiques. Dans le chapitre 7, nous avons découvert que si nous prenons en compte le biais *myside*, le nombre de personnes dans chaque camp au début d'une controverse est déterminant pour son issue. En outre, ce modèle a démontré que le filtrage des arguments faibles au niveau de la communauté pouvait modifier l'issue d'un débat en réduisant le nombre d'arguments faibles qui entrent dans le débat public, atténuant ainsi l'impact du biais *myside*.

Ces observations suggèrent que la compréhension des controverses va bien au-delà de la simple connaissance des arguments des deux parties : il est nécessaire de savoir non seulement quels arguments existent, mais aussi **qui a accès à quels arguments**. Il est donc nécessaire d'analyser les structures sociales qui influencent la manière dont les gens entrent en contact avec des arguments. Cette idée renforce la nécessité d'aborder les controverses comme des phénomènes à la fois *dynamiques*

et *sociaux*. Comme l'a démontré ce travail, les modèles basés sur des agents et les études empiriques sont des outils prometteurs et complémentaires pour explorer ces dimensions des controverses sociotechniques.

***Gish Gallop*, mauvaise foi et biais cognitifs**

En 1994, l'anthropologue Eugenie Scott a identifié une stratégie d'argumentation appelée "*Gish Gallop*" (Scott, 1994). Nommée d'après un éminent créationniste américain, Duane Gish, cette stratégie consiste à noyer son adversaire sous un déluge d'arguments qui peuvent être invalides, invérifiables, inexacts, délibérément trompeurs ou même hors de propos. Cette technique repose sur le fait que chaque argument sera beaucoup plus long à réfuter qu'à affirmer, un principe également connu sous le nom de loi de Brandolini (ou "principe d'asymétrie du *bullshit*", que nous avons décrit au chapitre 10). Ainsi, un opposant n'aura pas le temps de contrer tous les arguments produits par le *Gish Gallopper*, et le public, surtout s'il n'est pas familier avec le sujet et incapable d'évaluer les arguments lui-même, peut être fortement influencé.

Dans mon travail, j'ai identifié plusieurs situations, à la fois dans des simulations basées sur des agents et dans des études empiriques, qui ressemblent à des cas de *Gish Gallop*. Le chapitre 7 a démontré que des agents biaisés pouvaient remporter un débat et convaincre la communauté scientifique de leur position préférée, même lorsqu'il s'agissait de l'option la plus faible. Ce résultat n'était possible que lorsque les agents soutenant la position la plus faible étaient majoritaires au début du débat. La clé de cette dynamique réside dans le mécanisme de biais décrit par Mercier (2017), où le biais *myside* permet aux agents de générer plus facilement des arguments en faveur de leur position. Bien que ces arguments soient plus faibles, leur quantité submerge les partisans de la théorie la plus forte : incapables de les traiter tous, ils finissent par se laisser influencer. Ce mécanisme reflète la dynamique d'un *Gish Gallop*, à quelques différences près. La définition traditionnelle du *Gish Gallop* implique que deux adversaires débattent pour convaincre un public extérieur. En revanche, mon modèle décrit un débat au sein d'une communauté où les individus peuvent à la fois générer des arguments et changer d'opinion. Dans ce contexte, le nombre de personnes défendant l'option la plus faible devient très important, puisque la charge de l'argumentation est transférée à la minorité. Cela situe le concept de *Gish Gallop* dans un cadre plus social et distribué, où la dynamique de la taille du groupe et la production d'arguments influencent fortement l'issue des débats.

Dans l'étude des stratégies argumentatives des promoteurs et des opposants d'une scierie industrielle présentée au chapitre 9, j'ai identifié une stratégie qui s'apparente au *Gish Gallop*. L'entreprise promouvant le projet a avancé une multitude d'arguments relativement faibles. Il ne s'agissait pas de mensonges purs et simples, mais d'arguments triviaux et de promesses susceptibles d'être rompues à

l'avenir, telles que des assurances concernant les niveaux de pollution (peu probables au vu des antécédents de l'entreprise sur d'autres sites). Plutôt que les submerger de mensonges, cette tactique pourrait être mieux décrite comme une "vague d'arguments faibles". En submergeant l'opposition de nombreuses affirmations peu convaincantes mais plausibles, l'entreprise a renforcé l'impression donnée au public local par l'ensemble de ses preuves, tout en rendant difficile pour la partie adverse de contrer efficacement chaque argument.

L'exemple le plus clair de la stratégie de *Gish Gallop*, cependant, se trouve dans l'étude des arguments des opposants au climat présentée au chapitre 10. Les opposants au climat mobilisent efficacement les références scientifiques pour produire un grand nombre d'arguments, dont beaucoup sont auto-contradictoires ou manifestement faux. Cela reflète la tactique du *Gish Gallop* : en bombardant leur public d'un grand nombre d'arguments erronés ou trompeurs, ils créent le doute et rendent difficile pour leurs adversaires de réfuter chaque point individuellement. Comme dans le modèle décrit au chapitre 7, ce scénario implique plus de deux débatteurs : il existe un grand nombre de blogs et de sites web anticonformistes qui appliquent cette tactique simultanément (et, comme le montre Chavalarias et al. (2023), parfois de manière coordonnée). Le public est composé de profanes pour lesquels il est pratiquement impossible d'évaluer avec précision la force ou la validité des arguments scientifiques présentés, ce qui rend le *Gish Gallop* particulièrement efficace.

Ces différentes situations peuvent nécessiter différentes approches de modélisation. Dans le chapitre 9, j'ai proposé d'utiliser des **poids** pour les arguments pour tenir compte du fait que certains sont plus faibles que d'autres. L'utilisation d'une sémantique graduelle peut être pertinente pour modéliser des débats où les arguments les plus faibles peuvent s'accumuler et gagner du terrain s'ils sont suffisamment nombreux. En revanche, dans le chapitre 10, où nous avons identifié le mauvais usage des arguments scientifiques, je suggère de pondérer les relations d'attaque. Cela montre que le problème n'est pas simplement la qualité des arguments eux-mêmes, mais plutôt la **pertinence des attaques**. Dans ces cas, les opposants au climat interprètent ou représentent mal les preuves scientifiques. Plutôt que de décrire les arguments des opposants comme intrinsèquement faibles, il est plus approprié de remettre en question la validité de la relation de soutien qu'ils affirment entre les preuves scientifiques qu'ils fournissent - qui peuvent être parfaitement valables - et les affirmations qu'ils font. Ainsi, ces études de cas soulignent l'importance de modéliser à la fois la force des arguments et la force des relations d'attaque, afin de saisir toute la complexité de ces controverses sociotechniques.

Une autre distinction importante réside dans la *motivation* qui sous-tend ces stratégies. Dans le chapitre 10, l'auto-contradiction et le détournement délibéré des preuves scientifiques que nous avons observés témoignent de mauvaise foi. Les opposants au climat utilisent sciemment des tactiques comme sortir des recherches de leur contexte pour semer le doute. Leur stratégie implique une manipulation intentionnelle, ce qui devrait être reflété dans un modèle en tant que tactique de

désinformation délibérée. En revanche, dans le chapitre 7, les agents sont *inconscients* de leur partialité. Ils peuvent croire en la validité de leurs arguments, même s'ils sont influencés par des biais cognitifs qui les poussent à favoriser leur propre position. Leur comportement ne représente pas une tromperie, mais plutôt des caractéristiques cognitives qui conduisent à la production d'arguments biaisés. La situation dans le cas de la scierie (chapitre 9) est plus ambiguë. Il est difficile de savoir si les personels qui ont produit et présenté les arguments de l'entreprise sont partiaux ou s'ils sont de mauvaise foi. S'ils peuvent croire à leurs propres affirmations, il est tout aussi possible qu'ils produisent délibérément des arguments qu'ils savent faibles afin de protéger leurs intérêts. Cependant, qu'ils en soient conscients ou non, le résultat est le même : la partie adverse est submergée par une multitude d'arguments faibles.

L'existence de ces mécanismes de Gish Gallop met également en évidence le facteur crucial du *temps* dans les débats. En effet, le temps est une composante nécessaire de la stratégie de Gish Gallop : accabler son adversaire d'une succession rapide d'arguments faibles ou trompeurs fonctionne *précisément* parce que l'adversaire n'a pas assez de temps pour réfuter chacun d'entre eux. Cela crée un déséquilibre dans le débat, permettant au camp le plus faible de paraître plus fort simplement en raison de la quantité d'arguments produits. Cela se reflète dans le modèle à base d'agents du chapitre 7, où les agents minoritaires n'ont pas assez de temps pour réfuter les arguments produits par les nombreux agents majoritaires, qui peuvent influencer le débat malgré la qualité inférieure de leurs revendications. De la même manière que pour les structures sociales, nous constatons que les controverses sont bien plus qu'une simple carte d'arguments : elles évoluent dans le temps. Par conséquent, pour comprendre et modéliser pleinement ces controverses, il est essentiel de les représenter comme des processus dynamiques qui se développent et changent au fur et à mesure que le débat se déroule. Cette dimension temporelle est nécessaire pour rendre compte de stratégies telles que Gish Gallop, où la vitesse de production des arguments joue un rôle clé.

Le modèle présenté au chapitre 7 suggère une méthode potentielle pour contrer l'efficacité des stratégies de type Gish Gallop. Dans ce modèle, un filtre communautaire qui empêche les arguments faibles d'être partagés publiquement réduit l'avantage d'une majorité partielle. Ce filtre pourrait représenter des mécanismes tels que l'examen par les pairs dans la communauté scientifique, où les études de faible qualité ont moins de chances d'être publiées. Toutefois, la mise en œuvre d'un tel mécanisme dans des controverses sociotechniques plus larges, qui impliquent de multiples types d'acteurs, est beaucoup plus difficile. Le **Fact-checking** est un analogue potentiel de ce filtre dans le contexte des débats publics. Cependant, il a ses limites : il n'y a aucune garantie de l'indépendance des vérificateurs de faits, avec des cas possibles de partialité et de censure. En outre, l'analyse des stratégies de Gish Gallop dans des controverses réelles révèle une simplification critique de notre modèle : dans la simulation, le nombre d'arguments traités par le filtre est sans importance. Le filtre fonctionne de la même manière, quelle que soit la

quantité d'arguments produits. Cependant, dans la réalité, si une communauté scientifique ou des *fact-checkers* sont submergés par un flot d'arguments faibles, l'effort nécessaire pour les examiner, les réfuter ou les vérifier est considérable. Un modèle plus réaliste tiendrait compte des contraintes liées au filtrage de grandes quantités d'informations, en intégrant éventuellement des éléments tels que les ressources limitées et le temps dont disposent les *fact-checkers* ou les pairs.

Comment modéliser les controverses socio-techniques

En résumé, les études sur les controverses présentées dans ce manuscrit ont permis de tirer les enseignements suivants sur la manière de modéliser les controverses sociotechniques :

- **Utiliser des structures sociales** : La façon dont les arguments circulent et façonnent l'opinion des gens est fortement influencée par les structures sociales, telles que les réseaux d'acteurs et leurs canaux de communication. Les modèles doivent représenter ces structures et leur rôle dans la dynamique des controverses.
- **Importance de la dimension temporelle** : Les débats évoluent dans le temps, les stratégies et les arguments changeant au fur et à mesure que la controverse progresse. Cela nécessite des approches de modélisation dynamique telles que les modèles à base d'agents pour saisir la manière dont les arguments sont produits, communiqués et abandonnés au fil du temps. La nature évolutive des débats.
- **Inclure les biais et la désinformation délibérée** : Les controverses sociotechniques sont influencées non seulement par les biais cognitifs des participants, mais aussi par la désinformation délibérée. Cette double présence devrait être reflétée dans les modèles afin de tenir compte des distorsions involontaires et stratégiques de l'information.
- **Pondérer les arguments et les attaques** : Tous les arguments n'ont pas la même valeur et la pertinence ou la force des relations d'attaque peut varier. Les modèles doivent inclure des arguments et des attaques pondérés s'ils visent à refléter les préjugés et les stratégies d'argumentation de mauvaise foi.
- **Utiliser une sémantique graduelle** : Enfin, les études suggèrent que les modèles devraient utiliser une sémantique graduelle pour capturer l'impact cumulatif des arguments, qui est central dans les stratégies argumentatives que nous avons identifiées.

Informations concernant les outils

Cette section résume les réflexions sur ce que cette recherche peut nous apprendre sur les outils formels qui y ont contribué.

Argumentation abstraite

Mon travail illustre l'utilité et la flexibilité de l'argumentation abstraite. Premièrement, l'argumentation abstraite s'est avérée très utile pour modéliser la cognition des agents d'une manière simple mais informative. Deuxièmement, lorsqu'elle est utilisée dans le contexte d'études de cas réels, elle s'est avérée être un moyen fructueux de représenter la structure des controverses. Son utilisation pour décrire des controverses réelles nous permet d'identifier et de comprendre des caractéristiques et des stratégies argumentatives distinctes. En particulier, grâce aux outils formels concernant l'acceptabilité des arguments, nous avons pu identifier formellement l'auto-contradiction des opposants au climat dans le chapitre 10. Cela nous a permis d'étayer notre affirmation selon laquelle cette position n'est pas rationnelle, soulignant ainsi l'utilité des approches formelles dans l'analyse des études de cas.

Lorsque j'ai présenté des graphes d'argumentation abstraite aux participants du chapitre 9, j'ai remarqué que cela permettait une certaine réflexivité et une meilleure compréhension des deux positions du débat. Ces graphes peuvent contribuer à améliorer les stratégies d'argumentation des acteurs impliqués. Dans d'autres cas, ils peuvent aider à développer une compréhension de la position de la partie adverse, ce qui peut être utile pour parvenir à un consensus.

Comme développé ci-dessus, mon travail soutient l'utilisation d'approches pondérées et graduelles de l'argumentation abstraite, en accord avec de nombreux autres travaux empiriques. En conclusion, l'argumentation abstraite est un outil prometteur dans l'étude des controverses socio-techniques. En particulier, elle est fructueuse lorsqu'elle est combinée à des approches telles que les modèles multi-agents, ou lorsqu'elle est complétée par de la collecte de données.

Modèles multi-agents

Les modèles basés sur des agents nous permettent d'étudier les aspects dynamiques et sociaux des débats qui, comme nous l'avons montré, sont des composantes essentielles des controverses sociotechniques. Le travail de modélisation présenté dans la partie II a contribué à la compréhension de la complexité des controverses socio-techniques. Ces modèles ont mis en évidence l'importance des structures sociales, et la modélisation du biais *myside* en particulier a permis de mieux comprendre les phénomènes identifiés dans les études de cas. Comme développé ci-dessus, mon

travail identifie la présence de désinformation délibérée dans les controverses. A ce titre, il préconise la modélisation d'agents hétérogènes dont les motivations peuvent être distinctes.

Orientations futures de la recherche

Compléter le travail sur les études de cas

Les études de cas réels de controverses présentées dans les chapitres 9 et 10 sont des travaux en cours. La première orientation de mes recherches consistera à compléter ces travaux, en améliorant les graphes d'argumentation représentant chaque controverse. Le travail présenté dans le chapitre 9 sera complété en ajoutant des poids aux arguments et en appliquant une sémantique graduelle pour évaluer l'acceptabilité de chaque argument. Cela permettra de mieux comprendre les forces et les faiblesses de chaque position. Dans le chapitre 10, nous prévoyons de compléter la base de données des articles scientifiques mobilisés par les climato-sceptiques, en recueillant des informations supplémentaires sur le nombre d'arguments scientifiques qui soutiennent chacune des affirmations des opposants, afin de construire un cadre détaillé de la controverse.

Une chose à noter à propos de ces deux controverses est qu'elles sont d'échelles très différentes : le débat étudié au chapitre 9 implique principalement des acteurs locaux et est plutôt récent, alors que la controverse sur le changement climatique se déroule à un niveau international depuis de nombreuses années, impliquant des décennies de connaissances scientifiques et des centaines de milliers d'acteurs. Une étude plus approfondie des graphes d'argumentation pourrait nous permettre de comprendre comment l'échelle d'une controverse influence sa structure, sa dynamique et les stratégies d'argumentation déployées. À l'avenir, pour compléter ces analyses, j'aimerais étudier d'autres controverses, peut-être d'une échelle ou d'un thème différent, afin d'explorer les différences et les similitudes dans les structures, les stratégies ou les arguments.

Développement d'un modèle multi-agents de controverses sociotechniques

Une autre orientation future, plutôt naturelle, de cette recherche serait d'exploiter les connaissances acquises grâce à mes études empiriques pour créer un modèle basé sur des agents représentant des controverses socio-techniques plus larges. Inspiré des modèles de débats scientifiques que j'ai déjà créés, un tel modèle mettrait en œuvre les caractéristiques décrites ci-dessus. Il comporterait des pondérations pour les arguments et les attaques, et des agents hétérogènes, prenant en compte la présence de tentatives de manipulation avec des agents (coordonnés) débattant de mauvaise foi.

Cela nous permettrait d'explorer la dynamique induite par des stratégies délibérées de désinformation. Comme nous l'avons fait avec le filtre à biais au chapitre 7, nous pourrions également tester des moyens de les atténuer. Nous pourrions étudier l'interaction entre les biais cognitifs et la désinformation stratégique. Par exemple, nous pourrions étudier l'existence de boucles de rétroaction dans lesquelles les biais cognitifs et les phénomènes sociaux tels que l'homophilie renforcent l'efficacité de la désinformation.

Exploration des liens entre mon travail et la théorie de l'acteur-réseau

Un autre domaine de recherche prometteur consistera à explorer plus avant les liens significatifs entre l'argumentation abstraite et la théorie de l'acteur-réseau (ANT), comme nous avons commencé à le faire avec l'outil développé au chapitre 9. L'ANT est une approche bien étudiée et fructueuse pour l'analyse des controverses sociotechniques. En tant qu'informaticienne, ma formation ne m'a pas encore permis de m'engager pleinement dans ce corpus de travail, mais la comparaison des deux approches et des résultats serait du plus grand intérêt pour évaluer le pouvoir de l'argumentation abstraite dans l'étude des controverses. Nous pourrions vérifier si mes conclusions correspondent à celles obtenues avec l'ANT. Par exemple, dans le chapitre 9, j'ai noté que les participants qui pouvaient réfléchir à leurs propres stratégies argumentatives avaient tendance à modifier leurs positions. Cela rappelle l'un des principaux résultats de l'ANT, à savoir que les controverses, plutôt que de révéler des oppositions existantes, sont des situations qui conduisent à la création et au développement de camps opposés tout au long du conflit (Akrich et al., 2006). En outre, l'une des principales caractéristiques de l'ANT est qu'elle considère à la fois les humains et les non-humains - qui peuvent être des concepts, des technologies, des institutions ou des objets matériels - comme des acteurs qui influencent le réseau de relations au sein d'une controverse. Au contraire, les modèles de dynamique de l'opinion en général, et ceux qui utilisent l'argumentation abstraite en particulier, se concentrent généralement sur les agents humains et leurs interactions. Ainsi, en s'inspirant de la méthodologie ANT, nous pourrions améliorer nos modèles en incorporant des acteurs non humains lors de la modélisation des controverses sociotechniques. Cette intégration soulèverait des questions intéressantes, telles que la manière dont les acteurs non humains peuvent être représentés dans des graphes d'argumentation abstraite et la manière dont leur influence sur la circulation et la force des arguments peut être quantifiée. En outre, elle pourrait conduire au développement de nouveaux types de modèles qui rendent mieux compte de la complexité et de l'interconnexion des controverses dans le monde réel, en étendant l'applicabilité des modèles basés sur les agents et de la théorie de l'argumentation dans l'étude des controverses socio-technique.

Vulgariser et évaluer l'impact de ma recherche

Après avoir terminé mon doctorat, je travaillerai à la vulgarisation de certains de mes résultats auprès du public. Bien que cela ne fasse pas exactement partie de la recherche académique, la vulgarisation des résultats peut être extrêmement instructive et renforcer l'utilité de la recherche pour la société civile. En particulier, l'étude de la controverse entourant le projet de méga-scierie à Saint-Agnan a impliqué un lien avec le public local, ce qui rend sa communication très pertinente. Récemment, des militants écologistes locaux, intéressés par le projet de scierie mais ne le connaissant pas, m'ont contacté pour me demander un résumé des arguments des deux côtés de la controverse. Je me suis rendu compte que mon travail n'offrait pas encore un moyen simple de visualiser et de comprendre les positions des deux camps, en particulier pour les profanes peu familiarisés avec la programmation ou l'argumentation abstraite. Pourtant, l'intérêt de cette recherche et l'impact qu'elle pourrait avoir sur les connaissances des populations locales sont évidents. C'est pourquoi j'aimerais explorer les moyens de communiquer les controverses au public, éventuellement en utilisant l'argumentation abstraite. Par exemple, cela pourrait prendre la forme de cartes d'argumentation résumant la controverse de manière simple et visuelle. Cela permettrait également d'améliorer notre connaissance de l'utilité de l'argumentation abstraite dans la communication. Cela rappelle les travaux de [Vesic et al. \(2022\)](#), qui a étudié l'effet de la représentation visuelle des graphiques d'argumentation sur la cognition des gens.

Cet accent mis sur la vulgarisation s'inscrit dans un contexte plus large de réflexions personnelles sur l'utilisation de la science et le rôle qu'elle joue dans les profondes transformations auxquelles notre monde est confronté. A titre personnel, je pense que la crise environnementale globale à laquelle nous sommes confrontés, qui menace les conditions qui font de notre planète un environnement vivable pour nous et tous ses autres habitants, devrait être au cœur de nos efforts en tant que scientifiques. Ces questions m'ont amené à participer à plusieurs efforts collectifs de réflexion sur le rôle que la science (et en particulier l'informatique) joue, et pourrait jouer, dans la situation à laquelle nous sommes confrontés. Cette attention s'est reflétée dans le choix de mes sujets, et j'espère poursuivre dans cette voie, en explorant comment la science en général, et mon travail en particulier, peuvent être liés aux efforts déployés sur le terrain pour comprendre et combattre les causes de cette crise.

En résumé, le travail présenté dans ce manuscrit peut être considéré comme le début d'un projet de recherche plus large, utilisant l'argumentation abstraite de différentes manières pour explorer et comprendre les controverses socio-techniques. Étant donné la variété des questions que l'argumentation abstraite nous a déjà permis d'aborder, je suis convaincu que l'exploration plus poussée des questions de recherche exposées dans ce travail apportera de nouvelles perspectives et de nouveaux résultats intéressants. De plus, ce travail impliquera une interaction fascinante entre différents domaines d'étude, tels que l'argumentation formelle, les modèles basés sur les agents, la bibliométrie ou la théorie de l'acteur réseau. En même temps,

il contribuera à notre compréhension des questions actuelles urgentes soulevées par la désinformation scientifique.

Bibliography

- Abelson, R. P. (1964). Mathematical models of the distribution of attitudes under controversy. *Contributions to mathematical psychology*.
- Akers, R. L., Krohn, M. D., Lanza-Kaduce, L., and Radosevich, M. (1995). Social learning and deviant behavior: A specific test of a general theory. *Contemporary masters in criminology*, pages 187–214.
- Akrich, M., Callon, M., and Latour, B. (2006). *Sociologie de la traduction: textes fondateurs*. Presses des MINES.
- Amgoud, L. and Ben-Naim, J. (2018). Weighted bipolar argumentation graphs: Axioms and semantics. In *Twenty-Seventh International Joint Conference on Artificial Intelligence-IJCAI 2018*, pages 5194–5198.
- Amgoud, L., Ben-Naim, J., Doder, D., and Vesic, S. (2017). Acceptability semantics for weighted argumentation frameworks. In *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*. International Joint Conferences on Artificial Intelligence (IJCAI).
- AP-NORC (2024). Americans’ attitudes and actions on climate change. <https://epic.uchicago.edu/insights/2024-poll-americans-views-on-climate-change-and-policy-in-12-charts/>. Accessed: 2024-09-21.
- Arnott, D. (1998). A taxonomy of decision biases. *Monash University, School of Information Management and Systems, Caulfield*.
- Axelrod, R. (1997a). Advancing the art of simulation in the social sciences. In *Simulating social phenomena*, pages 21–40. Springer.
- Axelrod, R. (1997b). The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, 41(2):203–226.
- Aydinonat, N. E., Reijula, S., and Ylikoski, P. (2021). Argumentative landscapes: the function of models in social epistemology. *Synthese*, 199(1):369–395.
- Baccini, E., Christoff, Z., Hartmann, S., and Verbrugge, R. (2023). The wisdom of the small crowd: Myside bias and group discussion. *Journal of Artificial Societies and Social Simulation*, 26(4).
- Baccini, E. and Hartmann, S. (2022). The myside bias in argument evaluation: A bayesian model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.
- Bala, V. and Goyal, S. (1998). Learning from neighbours. *The review of economic studies*, 65(3):595–621.

- Banisch, S. and Shamon, H. (2021). Biased processing and opinion polarization: experimental refinement of argument communication theory in the context of the energy debate. *ArXiv*, abs/2212.10117.
- Banisch, S. and Shamon, H. (2022). Validating argument-based opinion dynamics with survey experiments.
- Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, 1(3):221–235.
- Baroni, P., Rago, A., and Toni, F. (2018). How many properties do we need for gradual argumentation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Baroni, P., Romano, M., Toni, F., Aurisicchio, M., and Bertanza, G. (2015). Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation*, 6(1):24–49.
- Barreteau, O., Bousquet, F., Étienne, M., Souchère, V., and d’Aquino, P. (2013). Companion modelling: a method of adaptive and participatory research. In *Companion Modelling: A participatory approach to support sustainable development*, pages 13–40. Springer.
- Benson, B. and Manoogian, J. (2017). Cognitive bias codex. <https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18>. Accessed: 2024-10-02.
- Besnard, P. and Hunter, A. (2001). A logic-based theory of deductive arguments. *Artificial Intelligence*, 128(1-2):203–235.
- Betz, G. (2022). Natural-language multi-agent simulations of argumentative opinion dynamics. *Journal of Artificial Societies and Social Simulation*, 25(1):nil.
- Bonzon, E. and Maudet, N. (2012). On the outcomes of multiparty persuasion. In *Argumentation in Multi-Agent Systems: 8th International Workshop, ArgMAS 2011, Taipei, Taiwan, May 3, 2011, Revised Selected Papers 8*, pages 86–101. Springer.
- Borg, A., Frey, D., Šešelja, D., and Straßer, C. (2017a). An argumentative agent-based model of scientific inquiry. In Benferhat, S., Tabia, K., and Ali, M., editors, *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I*, pages 507–510. Springer International Publishing, Cham.
- Borg, A., Frey, D., Šešelja, D., and Straßer, C. (2017b). Examining network effects in an argumentative agent-based model of scientific inquiry. In Baltag, A., Seligman, J., and Yamada, T., editors, *Logic, Rationality, and Interaction: 6th International Workshop, LORI 2017, Sapporo, Japan, September 11-14, 2017, Proceedings*, pages 391–406. Springer Berlin Heidelberg, Berlin, Heidelberg.

- Borg, A., Frey, D., Šešelja, D., and Straßer, C. (2018). Epistemic effects of scientific interaction: approaching the question with an argumentative agent-based model. *Historical Social Research*, 43(1):285–309.
- Borg, A., Frey, D., Šešelja, D., and Straßer, C. (2019). Theory-choice, transient diversity and the efficiency of scientific inquiry. *European Journal of Philosophy of Science*, (26).
- Bourdieu, P. (1973). L’opinion publique n’existe pas. *Questions de Sociologie*, 84.
- Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., and Holman, B. (2017a). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of science*, 84(1):115–159.
- Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., and Holman, B. (2017b). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, 84(1):115–159.
- Brock, W. A. and Durlauf, S. N. (1999). A formal model of theory choice in science. *Economic theory*, 14:113–130.
- Butler, G., Pigozzi, G., and Rouchier, J. (2019). Mixing dyadic and deliberative opinion dynamics in an agent-based model of group decision-making. *Complexity*, 2019:1–31.
- Caiati, C., Pollice, P., Favale, S., and Lepera, M. E. (2020). The herbicide glyphosate and its apparently controversial effect on human health: An updated clinical perspective. *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, 20(4):489–505.
- Carbon Brief (2011). Analysing the 900 papers supporting climate scepticism: Part 1, 2 and 3. Accessed: 2024-09-18.
- Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591–646.
- Catellani, A. and Lits, G. (2024). *Transitions en tension: Controverses et tensions autour des transitions écologiques*, volume 3. ISTE Group.
- Cayrol, C. and Lagasque-Schiex, M.-C. (2005). On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 8th European Conference, ECSQARU 2005, Barcelona, Spain, July 6-8, 2005. Proceedings 8*, pages 378–389. Springer.
- Centola, D., Gonzalez-Avella, J. C., Eguiluz, V. M., and San Miguel, M. (2007). Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution*, 51(6):905–929.
- Cerutti, F., Cramer, M., Guillaume, M., Hadoux, E., Hunter, A., and Polberg, S. (2021). Empirical cognitive studies about formal argumentation.

- Chateauraynaud, F., Reber, B., and Van Meter, K. (2003). Marlowe, prospero & la technologie littéraire. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 79(1):5–46.
- Chavalarias, D., Bouchaud, P., Chomel, V., and Panahi, M. (2023). Les nouveaux fronts du déniisme et du climato-scepticisme.
- Clark, C. and Tetlock, P. (2021). *Adversarial Collaboration: The Next Science Reform*.
- Clifford, P. and Sudbury, A. (1973). A model for spatial conflict. *Biometrika*, 60(3):581–588.
- Coan, T. G., Boussalis, C., Cook, J., and Nanko, M. O. (2021). Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.
- Cocarascu, O., Rago, A., and Toni, F. (2019). Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *AAMAS*, pages 1261–1269.
- Cramer, M. and Guillaume, M. (2018). Empirical cognitive study on abstract argumentation semantics. In *Computational Models of Argument*, pages 413–424. IOS Press.
- Cramer, M. and Guillaume, M. (2019). Empirical study on human evaluation of complex argumentation frameworks. In *Logics in Artificial Intelligence: 16th European Conference, JELIA 2019, Rende, Italy, May 7–11, 2019, Proceedings 16*, pages 102–115. Springer.
- Crawford, M. B. (2009). *Shop class as soulcraft: An inquiry into the value of work*. Penguin.
- d’Allens, G. (2019). *Main basse sur nos forêts*. Média Diffusion.
- Deffuant, G., Amblard, F., Weisbuch, G., and Faure, T. (2002). How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4).
- Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98.
- Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.
- Dupuis de Tarlé, L. (2023). A gradual semantic to model opinion from bipolar argumentation graphs. In *Proceedings of the 16èmes Journées d’Intelligence Artificielle Fondamentale (JIAF 2023)*.

- Dupuis de Tarlé, L., Bonzon, E., and Maudet, N. (2022a). Multiagent dynamics of gradual argumentation semantics. In *21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*.
- Dupuis de Tarlé, L., Michelini, M., Borg, A., Pigozzi, G., Rouchier, J., Šešelja, D., and Straßer, C. (2024). An agent-based model of myside bias in scientific debates. *Journal of Artificial Societies & Social Simulation*, 27(3).
- Dupuis de Tarlé, L., Pigozzi, G., and Rouchier, J. (2022b). Spécificités de l'argumentation scientifique dans un débat. In *Proceedings of the 30èmes Journées Francophones sur les Systèmes Multi-Agents (JFSMA 2022)*, pages 53–62.
- Dupuis de Tarlé, L., Pigozzi, G., and Rouchier, J. (2023). Exploring opinion diversity and epistemic success with an argumentative model. In *Advances in Social Simulation: Proceedings of the 18th Social Simulation Conference, Glasgow, UK, 4–8 September 2023*, page 383. Springer.
- Dupuis de Tarlé, L., Michellini, M., Šešelja, D., and Straßer, C. (Forthcoming). Agent based models of argumentation. *The Handbook of Formal Argumentation Volume 3*.
- Dutilh Novaes, C. (2018). The enduring enigma of reason. *Mind & Language*, 33(5):513–524.
- Edmonds, B. and Moss, S. (2004). From kiss to kids—an ‘anti-simplistic’ modelling approach. In *International workshop on multi-agent systems and agent-based simulation*, pages 130–144. Springer.
- Ellis, G. (2018). So, what are cognitive biases? *Cognitive biases in visualizations*, pages 1–10.
- Eos (2020). Don't @ me: What happened when climate skeptics misused my work. <https://eos.org/opinions/dont-at-me-what-happened-when-climate-skeptics-misused-my-work>. Accessed: 2024-09-20.
- Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. math. inst. hung. acad. sci.*, 5(1):17–60.
- Ferrand, A. (2016). *Appartenances multiples, opinion plurielle*. Presses universitaires du Septentrion.
- Flache, A. and Macy, M. W. (2007). Local convergence and global diversity: the robustness of cultural homophily. *arXiv preprint physics/0701333*.
- Fu, G. and Zhang, W. (2016). Opinion formation and bi-polarization with biased assimilation and homophily. *Physica A: Statistical Mechanics and its Applications*, 444:700–712.
- Fu, H.-Z. and Waltman, L. (2022). A large-scale bibliometric analysis of global climate change research between 2001 and 2018. *Climatic Change*, 170(3):36.

- Gabbriellini, S. and Torroni, P. (2014). A new framework for abms based on argumentative reasoning. In *Advances in Intelligent Systems and Computing*, Advances in Intelligent Systems and Computing, pages 25–36. Springer Berlin Heidelberg.
- Gabriel, N. and O'Connor, C. (2022). Can confirmation bias improve group learning?
- Gardner, M. (1970). The fantastic combinations of john conway's new solitaire game'life. *Sc. Am.*, 223:20–123.
- Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research Online*, 2(2):1–15.
- Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.
- Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.
- Grisoni, A. (2015). Les mouvements d'opposition contre les grands projets inutiles imposés (gpil) à l'avant-garde de la ruralité? le mouvement notav contre le lyonnais. *Mouvements*, (4):126–130.
- GWPF (2024). Global warming policy foundation homepage. <https://www.thegwpf.org/>. Accessed: 2024-09-19.
- Hahn, U. and Hornikx, J. (2016). A normative framework for argument quality: argumentation schemes with a bayesian foundation. *Synthese*, 193(6):1833–1873.
- Haider, J., Söderström, K. R., Ekström, B., and Rödl, M. (2024). Gpt-fabricated scientific papers on google scholar: Key features, spread, and implications for pre-empting evidence manipulation. *Harvard Kennedy School Misinformation Review*, 5(5):1–16.
- Hales, D., Rouchier, J., and Edmonds, B. (2003). Model-to-model analysis. *Journal of Artificial Societies and Social Simulation*, 6(4).
- Haselton, M. G., Nettle, D., and Andrews, P. W. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 724–746.
- Hegselmann, R. and Krause, U. (2002). Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation (JASSS)*, 5(3).
- Hegselmann, R., Krause, U., et al. (2006). Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, 9(3):10.
- Hermansson, H. (2007). The ethics of nimby conflicts. *Ethical theory and moral practice*, 10:23–34.
- Holley, R. A. and Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663.

- Hong, L. and Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.
- Hunter, A. (2013). A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81.
- Jankó, F., Móricz, N., and Vancsó, J. P. (2014). Reviewing the climate change reviewers: Exploring controversy through report references and citations. *Geoforum*, 56:17–34.
- Jankó, F., Papp Vancsó, J., and Móricz, N. (2017). Is climate change controversy good for science? ipcc and contrarian reports in the light of bibliometrics. *Scientometrics*, 112(3):1745–1759.
- Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy*, 87(1):5–22.
- Kopecky, F. (2022). Arguments as drivers of issue polarisation in debates among artificial agents. *Journal of Artificial Societies and Social Simulation*, 25(1).
- Kudsk, P. and Mathiassen, S. K. (2020). Pesticide regulation in the european union and the glyphosate controversy. *Weed Science*, 68(3):214–222.
- Kuhn, T. (2000). *The Road since Structure*. University of Chicago Press, Chicago.
- Latour, B. (1999). On recalling ant. *The sociological review*, 47(1_suppl):15–25.
- Latour, B. and Woolgar, S. (1979a). *Laboratory life: The construction of scientific facts*. Princeton university press.
- Latour, Bruno, C. and Woolgar, S. (1979b). *Laboratory Life: the Construction of Scientific Facts*. Princeton: Princeton University Press.
- Ledoyen, F., Thomopoulos, R., Couture, S., Sadou, L., and Taillandier, P. (2021). An agent-based model representing the exchanges of arguments to accurately simulate the process of innovation diffusion. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, page nil.
- Lemieux, C. (2007). À quoi sert l’analyse des controverses? *Mil neuf cent*, (1):191–212.
- Liu, Q., Zhao, J., and Wang, X. (2015). Multi-agent model of group polarisation with biased assimilation of arguments. *IET Control Theory & Applications*, 9(3):485–492.
- Longino, H. (2002). *The Fate of Knowledge*. Princeton University Press, Princeton.
- Marx, W., Haunschild, R., Thor, A., and Bornmann, L. (2017). Which early works are cited most frequently in climate change research literature? a bibliometric approach based on reference publication year spectroscopy. *Scientometrics*, 110:335–353.

- Marzagão, D. K., Murphy, J., Young, A. P., Gauy, M. M., Luck, M., McBurney, P., and Black, E. (2018). Team persuasion. In *Theory and Applications of Formal Argumentation*, Theory and Applications of Formal Argumentation, pages 159–174. Springer International Publishing.
- Mäs, M. and Bischofberger, L. (2015). Will the personalization of online social networks foster opinion polarization? *Available at SSRN 2553436*.
- Mäs, M. and Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PLoS ONE*, 8(11):e74516.
- Mäs, M., Flache, A., Takács, K., and Jehn, K. A. (2013). In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. *Organization science*, 24(3):716–736.
- Mayo-Wilson, C. and Zollman, K. J. (2021). The computational philosophy: simulation as a core philosophical method. *Synthese*, 199(1):3647–3673.
- Mercier, H. (2016). Confirmation bias–myside bias. In *Cognitive illusions*, pages 109–124. Psychology Press.
- Mercier, H. (2017). Confirmation bias–myside bias. In Pohl, R. F., editor, *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory*, page 99–114. Routledge/Taylor & Francis Group.
- Mercier, H., Bonnier, P., and Trouche, E. (2016). 10 why don’t people produce better arguments? *Cognitive unconscious and human rationality*, page 205.
- Mercier, H. and Heintz, C. (2014). Scientists’ argumentative reasoning. *Topoi*, 33(2):513–524.
- Mercier, H. and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.
- Mercier, H. and Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.
- Mobilia, M. (2003). Does a single zealot affect an infinite group of voters? *Physical review letters*, 91(2):028701.
- Mossakowski, T. and Neuhaus, F. (2018). Modular semantics and characteristics for bipolar weighted argumentation graphs. *arXiv preprint arXiv:1807.06685*.
- Mutz, R., Bornmann, L., and Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: a general estimating equations approach. *PLoS ONE*, 7(10).
- No Tricks Zone (2024). No tricks zone homepage. <https://notrickszone.com/>. Accessed: 2024-09-20.

- No Tricks Zone - Sun Drives Climate (2024). 100+ papers – sun drives climate. <https://notrickszone.com/100-papers-sun-drives-climate/>. Accessed: 2024-09-20.
- O'Connor, C. and Weatherall, J. O. (2019). *The misinformation age: How false beliefs spread*. Yale University Press.
- Oreskes, N. and Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.
- O'Connor, C. and Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, 8:855–875.
- Pera, M. (1994). *The Discourses of Science*. The University of Chicago Press, Chicago, London.
- Perkins, D. (2019). Learning to reason: the influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues. *Judgment and Decision Making*, 14(6):624–643. Publication of the original unpublished note from 1986.
- Peters, U. (2020). Illegitimate values, confirmation bias, and mandevillian cognition in science. *The British Journal for the Philosophy of Science*.
- Polberg, S. and Hunter, A. (2018). Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *International Journal of Approximate Reasoning*, 93:487–543.
- Popper, K. R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, England: Routledge.
- Potyka, N. (2018). Continuous dynamical systems for weighted bipolar argumentation. In *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*.
- Potyka, N. (2019). Open-mindedness of gradual argumentation semantics. In *Scalable Uncertainty Management: 13th International Conference, SUM 2019, Compiègne, France, December 16–18, 2019, Proceedings 13*, pages 236–249. Springer.
- Proietti, C. and Chiarella, D. (2023). The role of argument strength and informational biases in polarization and bi-polarization effects. *Journal of Artificial Societies and Social Simulation*, 26(2):5.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Rago, A., Baroni, P., and Toni, F. (2018). On instantiating generalised properties of gradual argumentation frameworks. In *International Conference on Scalable Uncertainty Management*, pages 243–259. Springer.

- Rago, A., Li, H., and Toni, F. (2023). Interactive explanations by conflict resolution via argumentative exchanges. *(forthcoming) 20th International Conference on Principles of Knowledge Representation and Reasoning*.
- Rago, A., Toni, F., Aurisicchio, M., and Baroni, P. (2016). Discontinuity-free decision support with quantitative argumentation debates. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- Rahwan, I., Madakkatel, M. I., Bonnefon, J.-F., Awan, R. N., and Abdallah, S. (2010). Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, 34(8):1483–1502.
- Rescher, N. (2007). *Dialectics*. De Gruyter.
- Rosenstock, S., Bruner, J., and O'Connor, C. (2017). In epistemic networks, is less really more? *Philosophy of Science*, 84(2):234–252.
- Rouchier, J. (2024). L'éthique des modèles de simulation : une question d'interprétation ? *L'Anne sociologique*, 74:495–527.
- Salliou, N. and Thomopoulos, R. (2018). Modelling multicriteria argument networks about reduced meat consumption. In *FOODSIM 2018-10th International Conference on Simulation and Modelling in the Food and Bio-industry*, pages 1–6.
- Sanfeliu, A. and Fu, K.-S. (1983). A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (3):353–362.
- Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.
- Scott, E. (1994). Debates and the globetrotters. *The TalkOrigins Archive*.
- Šešelja, D. (2022). What kind of explanations do we get from agent-based models of scientific inquiry?
- Šešelja, D. and Straßer, C. (2013). Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190:2195–2217.
- Šešelja, D. and Weber, E. (2012). Rationality and irrationality in the history of continental drift: Was the hypothesis of continental drift worthy of pursuit? *Studies in History and Philosophy of Science*, 43:147–159.
- Sharman, A. (2014). Mapping the climate sceptical blogosphere. *Global Environmental Change*, 26:159–170.
- Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranganani, A., and Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 176(9):2243–2269.

- Singer, D. J., Bramson, A. L., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A., and Berger, W. J. (2018). Rational social and political polarization. *Philosophical Studies*, pages 1–25.
- Smart, P. R. (2018). Mandevillian intelligence. *Synthese*, 195(9):4169–4200.
- Squazzoni, F. (2009). *Epistemological aspects of computer simulation in the social sciences*, volume 5466. Springer.
- Stanovich, K. E., West, R. F., and Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264.
- Sulik, J., Bahrami, B., and Deroy, O. (2022). The diversity gap: when diversity matters for knowledge. *Perspectives on Psychological Science*, 17(3):752–767.
- Taillandier, P., Salliou, N., and Thomopoulos, R. (2021). Introducing the argumentation framework within agent-based models to better simulate agents’ cognition in opinion dynamics: Application to vegetarian diet diffusion. *Journal of Artificial Societies and Social Simulation*, 24(2):6.
- Thiriot, S. (2020). Small world is not enough: criteria for network choice and conclusiveness of simulations. *arXiv preprint arXiv:2002.03717*.
- Troupe, M. (2018). *The Zad and NoTAV: Territorial Struggles and the Making of a New Political Intelligence*. Verso Books.
- Vazquez, F., Krapivsky, P. L., and Redner, S. (2003). Constrained opinion dynamics: Freezing and slow evolution. *Journal of Physics A: Mathematical and General*, 36(3):L61.
- Vesic, S., Yun, B., and Teovanovic, P. (2022). Graphical representation enhances human compliance with principles for graded argumentation semantics. In *AA-MAS*.
- Wagener, A. (2024). Controverses discursives autour des véhicules électriques en france : entre transition écologique et doute. In *Transitions en tension: Controverses et tensions autour des transitions écologiques*, chapter 10. ISTE Group.
- Weatherall, J. O., O’Connor, C., and Bruner, J. P. (2020). How to beat science and influence people: Policymakers and propaganda in epistemic networks. *The British Journal for the Philosophy of Science*.
- Wolfe, C. R. and Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & reasoning*, 14(1):1–27.
- Wolfram, S. and Gad-el Hak, M. (2003). A new kind of science. *Appl. Mech. Rev.*, 56(2):B18–B19.

- Xu, Q., Huet, S., Perret, E., and Deffuant, G. (2020). Do farm characteristics or social dynamics explain the conversion to organic farming by dairy farmers? an agent-based model of dairy farming in 27 french cantons. *Journal of Artificial Societies and Social Simulation*, 23(2).
- Zamora Bonilla, J. (1999). The elementary economics of scientific consensus. *Theoria: An International Journal for Theory, History and Foundations of Science*, pages 461–488.
- Zhong, J. and Negre, E. (2022). A 3 r: Argumentative explanations for recommendations. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE.
- Zollman, K. (2007). The communication structure of epistemic communities. *Philosophy of science*, 74(5):574–587.
- Zollman, K. (2011). Computer simulation and emergent reliability in science. *Journal of Artificial Societies and Social Simulation*, 14(4):15.

Appendix

A Appendix of Chapter 5

In this section, I provide the proofs for Property 2 of Chapter 5, showing that the LSS semantics verifies open-mindedness and duality.

In the following, we will use σ_L to designate the LSS semantics, and ι_L and α_L respectively its influence and aggregation function.

A.1 Open-Mindedness

We recall the definition of open-mindedness.

Definition 31 (Open Mindedness). *Let $\sigma : G \times A \rightarrow D$ a semantic for weighted bipolar graphs on an interval D . The semantic is open-minded for all graph $B = \langle A, R, S, W \rangle$ if for all argument $a \in A$ and all $\epsilon > 0$, the following condition is satisfied : there exists a number $N \in \mathbb{N}$ such that if we add N new arguments whose base score is maximal : $A_N = a_1, \dots, a_N$, $A \cap A_N = \emptyset$, then :*

1. *For graph $B_{att} = \langle A \cup A_N, R \cup \{(a_i, a) | 1 \leq i \leq N\}, S, W' \rangle$, we obtain $|\sigma(B_{att}, a) - \min(D)| < \epsilon$*
2. *For graph $B_{supp} = \langle A \cup A_N, S \cup \{(a_i, a) | 1 \leq i \leq N\}, S, W' \rangle$, we obtain $|\sigma(B_{supp}, a) - \max(D)| < \epsilon$*

where $W'(b) = W(b)$ for all $b \in A$ and $W'(a_i) = \max(D)$ for $i \in \llbracket 1, N \rrbracket$.

For an acyclic bipolar graph $B = \langle A, R, S, W \rangle$, the LSS semantics is defined as the composition of an influence function ι_L and an aggregation function α_L . The influence function is defined as:

$$\iota_L : x \longrightarrow 1 - \frac{1}{1 + \epsilon^x}$$

This function has an image interval of $]0, 1[$, and is monotonously increasing, 0 being the limit when x tends to $-\infty$ and 1 the limit when x tends to ∞ .

Let $B = \langle A, R, S, W \rangle$ be an acyclic bipolar graph, $a \in A$ an argument such that $w_a \in]0, 1[$ and $\epsilon > 0$. In order to show that the LBS semantic is open-minded, we

must prove that we can bring the score of an argument as close to 0 as we want by adding attacking arguments, and as close to 1 as we want by adding supports. We only detail the reasoning for the attack case, the support one being similar.

Let us define $M = \ln \frac{1}{1-\epsilon} - 1$.

Case 1: $\alpha_L(B, a) = E(B, a) + c(w_a) < M$

We can add $N = 0$ new arguments and build $B_{att} = \langle A \cup A_N, R \cup \{(a_i, a) | 1 \leq i \leq N\}, S, W' \rangle = B$. Then we have $E(B_{att}, a) + c(w_a) = E(B, a) + c(w_a) = \alpha(B, a)$ and:

$$\begin{aligned} \alpha_L(B, a) &< M \\ \alpha_L(B_{att}, a) &< \ln \frac{1}{1-\epsilon} - 1 \\ \iota_L(\alpha_L(B_{att}, a)) &< \epsilon \\ \sigma_L(B_{att}, a) &< \epsilon \end{aligned}$$

Case 2: $\alpha_L(B, a) = E(B, a) + c(w_a) \geq M$

Let us take $N \in \mathbb{N}$ such that $N > E(B, a) + b(w_a) - M$. Then, we build B_{att} by adding N new arguments of base score 1 which attack a : $B_{att} = \langle A \cup A_N, R \cup \{(a_i, a) | 1 \leq i \leq N\}, S, W' \rangle$. Thus, we have:

$$\begin{aligned} E(B, a) + c(w_a) - N &< M \\ E(B_{att}, a) + c(w_a) &< M \\ \iota_L(\alpha_L(B_{att}, a)) &< \epsilon \end{aligned}$$

A.2 Duality

Duality is defined as follows.

Definition 32 (Duality). *Let $\sigma : G \times A \rightarrow D$ be a modular semantic for weighted bipolar graphs, with α its aggregation function and ι its influence function. The semantic σ verifies **duality** for all $B = \langle A, R, S, W \rangle$ if and only if it verifies the following property:*

Let $a, b \in A$ such that $w_a = 0.5 + \epsilon$, $w_b = 0.5 - \epsilon$ for an $\epsilon \in [0, 0.5[$, and the supporters and attackers of a and b are such that :

$$\begin{aligned} \alpha(B \setminus Att(B, a), a) &= -\alpha(B \setminus Supp(B, b), b) \\ \alpha(B \setminus Att(B, b), b) &= -\alpha(B \setminus Supp(B, a), a) \end{aligned}$$

then $\sigma(B, a) - w_a = w_b - \sigma(B, b)$.

Let us assume $B = \langle A, R, S, W \rangle$ an acyclic bipolar graph and $a, b \in A$ such that $w_a = 0.5 + \epsilon, w_b = 0.5 - \epsilon$ for an $\epsilon \in [0, 0.5]$, and the supporters and attackers of a and b are such that :

$$\begin{aligned}\alpha(B \setminus Att(B, a), a) &= -\alpha(B \setminus Supp(B, b), b) \\ \alpha(B \setminus Att(B, b), b) &= -\alpha(B \setminus Supp(B, a), a)\end{aligned}$$

We will show that $\sigma_{LSS}(B, a) + \sigma_{LSS}(B, b) = w_a + w_b = 1$.

First, the conditions on the attackers of b and a gives us that:

$$\begin{cases} \sum_{x \in Supp(B, a)} \sigma_{LSS}(x) + b(w_a) &= \sum_{x \in Att(B, b)} \sigma_{LSS}(x) + b(w_b) \\ \sum_{x \in Supp(B, b)} \sigma_{LSS}(x) + b(w_b) &= \sum_{x \in Att(B, a)} \sigma_{LSS}(x) + b(w_a) \end{cases}$$

By combining these two equations, we obtain:

$$\begin{aligned} \sum_{x \in Supp(B, a)} \sigma_L(x) + b(w_a) - \sum_{x \in Att(B, a)} \sigma_L(x) - b(w_a) &= \sum_{x \in Att(B, b)} \sigma_L(x) + b(w_b) - \sum_{x \in Supp(B, b)} \sigma_L(x) - b(w_b) \\ E(B, a) &= -E(B, b) \end{aligned}$$

Thus, we have that:

$$\begin{aligned} \alpha_L(B, a) + \alpha_L(B, b) &= E(B, a) + b(w_a) + E(B, b) + b(w_b) \\ &= b(w_a) + b(w_b) \\ &= \ln\left(\frac{1}{0.5 - \epsilon} - 1\right) + \ln\left(\frac{1}{0.5 + \epsilon} - 1\right) \\ &= \ln\left[\left(\frac{1}{0.5 - \epsilon} - 1\right)\left(\frac{1}{0.5 + \epsilon} - 1\right)\right] \\ &= \ln\left(\frac{1}{(0.5 - \epsilon)(0.5 - \epsilon)} - \frac{1}{0.5 - \epsilon} - \frac{1}{0.5 + \epsilon} + 1\right) \\ &= \ln\left(1 + \frac{1 - 0.5 + \epsilon - 0.5 + \epsilon}{(0.5 + \epsilon)(0.5 - \epsilon)}\right) \\ &= \ln(1) \\ &= 0 \end{aligned}$$

Finally:

$$\begin{aligned}
\sigma_L(B, a) + \sigma_L(B, b) &= \iota_L(\alpha_L(B, a)) + \iota_L(\alpha_L(B, b)) \\
&= 1 - \frac{1}{1 + e^{\alpha(B, a)}} + 1 - \frac{1}{1 + e^{\alpha(B, b)}} \\
&= 2 - \left(\frac{1 + e^{\alpha(B, b)} + 1 + e^{\alpha(B, a)}}{(1 + e^{\alpha(B, a)})(1 + e^{\alpha(B, b)})} \right) \\
&= 2 - \left(\frac{2 + e^{\alpha(B, b)} + e^{\alpha(B, a)}}{1 + e^{\alpha(B, b)} + e^{\alpha(B, a)} + \underbrace{e^{\alpha(B, a) + \alpha(B, b)}}_{=1}} \right) \\
&= 2 - 1 \\
&= 1
\end{aligned}$$

B Appendix of Chapter 7

In this section, I provide additional details and results from the model of Chapter 7, including a proof of Observation 1.

B.1 Algorithm for computing the acceptable arguments

This algorithm for generating the grounded labelling starts by selecting all arguments that are not attacked, and then iteratively: any argument that is attacked by an argument that has just been selected in the extension is considered unacceptable, and then we add to the extension the arguments all of whose attackers are unacceptable.

In our setting, the arguments which are not attacked are the leafs of each CAAT.

Algorithm 1 Algorithm to compute the acceptable arguments of C

C is a CAAT.

$G \leftarrow \text{leaf}(C)$ $O \leftarrow$ empty list
 while length of O + length of G < length of C do
 n in G $O \leftarrow O +$ successors of n o in O a in successors of o a not in O and predecessors of a is included in O $G \leftarrow G + \{a\}$

B.2 Additional results

Table 1 register how different parameters of *DebSet* affect the coefficient b of the function

$$\overline{Supp} = \frac{\alpha}{1 + e^{-\beta \cdot d}}. \quad (1)$$

As mentioned in the main article, the if β increases, the steepness of the curve also increases, which makes the community more sensitive to changes in the area around $d = 0$ and less sensitive to changes in most distant areas.

Table 1: The impact on β is reported, when the parameter on the left column increases.

Parameter	Impact on β
p_{see}	β increases.
N_{CA}	β increases.
$StDev$	β decreases.

B.3 Proof of Observation 1

To substantiate the (quite surprising) computational findings about biased, equally distributed communities, we provide an analytical result about such populations. To do so, we restrict our analysis to a simple case, where we show that our computational results are corroborated. Indeed, as the number of steps increases, the number of possible states of the model grows rapidly, which renders a mathematical analysis very challenging.

We compare two communities, a biased and a non biased one. Each community is composed of two agents, where one agent prefers $ResProg_1$ and the other prefers $ResProg_2$. Crucially, we find that after one round of simulation the probability for each possible outcome of the agents (e.g. the two of them preferring $ResProg_1$) are the same, whether or not the agents are biased. Formally, this amounts to the following.

Observation 2. *Let $Com = \langle N, n_1^0, p_{see}, bias \rangle$ and $Com' = \langle N, n_1^0, p_{see}, bias' \rangle$ be two communities with $N = 2, p_{see} = 1, n_1^0 = 1, bias = 0$ and $bias' > 0$. For any environment Env, S_1, S_2 such that $S_1, S_2 \in]bias', 1 - bias'[$ and $N_{CA} = 1$, the expected value of the support n_1^t at time $t = 1$ (with possible outcomes $\{0, 1, 2\}$) is the same for Com and Com' .*

If confronted with the same environment Env , two equally distributed communities of two agents have the same expected value of support n_1^t at time $t = 1$ even if one community is biased and the other is not.

Proof. First we shall specify more formally the notion of expected value. Given a debate setting $DebSet$ with $|N| = 2$, it is possible to define an event space $\omega_n = \{0, 1, 2\}$ which contains every possible outcome for the variable n_1^1 . In addition, based on how each action of the agents is defined in the protocol, it is possible

to assign to each of these events a probability $P : \omega_n \mapsto [0, 1]$. Consequently, when talking about expected value $\mathbb{E}$ for n_1^1 , we consider it to be the following: $\mathbb{E}(n_1^1) = P(n_1^1 = 2) \cdot 2 + P(n_1^1 = 1) \cdot 1$. Notably, the probability $P(n_1^1 = i)$ can be computed by computing the probability of even more elementary events, e.g. the event that one of the two agents decides to attack one argument. We consider an environment where $N_{CA} = 1$, $StDev = 0.2$ and two communities Com and Com' where $N = N'$, $n_1^0 = n_1^{0'}$, $p_{see} = p'_{see} = 1$ and $0 = bias < bias'$.

By definition, $\mathbb{E}(n_1^1) = P(n_1^1 = 2) \cdot 2 + P(n_1^1 = 1) \cdot 1$, and, similarly, for $\mathbb{E}(n_1^{1'})$. We shall denote with A the agent who starts with a preference for $ResProg_1$ and B the one who starts preferring $ResProg_2$. We denote with C_1 and C_2 the central arguments of $ResProg_1$ and $ResProg_2$ respectively. During the first step of the debate, each agent can either attack C_1 , attack C_2 or produce no attack. We define the event $Att(C_i) = k$ as the event in which after the first step, the number of attacks against C_i is k .

For an agent to change her mind, the score of the other $ResProg$ must be strictly superior to the score of her currently preferred $ResProg$. Because we consider $p_{see} = 1$ and thus the agents are aware of all the attacks which are produced, we have the following inequalities :

- $P(n_1^1 = 2) = P((Att(C_2) = 2 \cap Att(C_1) = 0) \cup (P(Att(C_2) = 1 \cap Att(C_1) = 0)))$
- $P(n_1^1 = 1) = (P(Att(C_2) = 0 \cap Att(C_1) = 0)) \cup (P(Att(C_2) = 1 \cap Att(C_1) = 1))$

The same holds for when we talk about a biased community. To show when we are considering a biased community, we use the notation P_{bias} , e.g. $P_{bias}(Att(C_2) = 2 \cap Att(C_1) = 0)$ refers to the probability of obtaining to attacks against C_1 and none against C_1 with a biased community.

We proceed as follows.

1. We prove, extensively, that $P(Att(C_2) = 1, Att(C_1) = 0) = P_{bias}(Att(C_2) = 1, Att(C_1) = 0)$, i.e. that both the biased community and the unbiased one have the same probability of producing an argument against C_2 .
2. Thus, we mention, briefly that following the same steps as above the following can be proven:

- (a) that $P(Att(C_2) = 0, Att(C_1) = 0) = P_{bias}(Att(C_2) = 0, Att(C_1) = 0)$,
- (b) $P(Att(C_2) = 2, Att(C_1) = 0) = P_{bias}(Att(C_2) = 2, Att(C_1) = 0) + \frac{bias^2}{4}$,
and that
- (c) $P(Att(C_2) = 1, Att(C_1) = 1) = P_{bias}(Att(C_2) = 1, Att(C_1) = 1) - \frac{bias^2}{2}$.

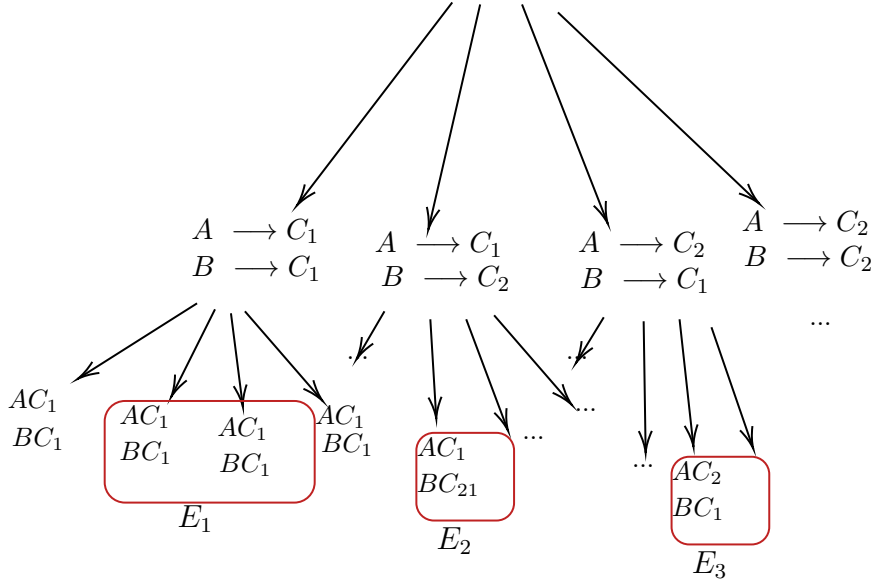


Figure 2: Representation of the probability tree.

3. Finally, we combine every statement together and we prove the initial observation.

Through the first point, we hope to give the reader a feeling of how all the other computations should be done, and we then leave them out in the second point.

Claim 1. We shall prove that $P(Att(C_2) = 1, Att(C_1) = 0) = P_{bias}(Att(C_2) = 1, Att(C_1) = 0)$. In short, the probability of the first step ending with one attack on C_1 and none on C_2 is the same for both agents. Figure 2 presents a schematic view of the probability tree associated with this problem. We use the following names of events.

- A (resp. B) $\rightarrow C_i$: Agent A (resp. B) is investigating argument C_i .
- A (resp. B) C_i : Agent A (resp. B) has produced an attack against C_i .
- A (resp. B) $\neg C_i$: Agent A (resp. B) has failed to produce an attack against C_i .

We shall start with the case for community Com , where $bias = 0$. First, we notice that the event $E = Att(C_1) = 1 \cap Att(C_2) = 0$ is the union of three events circled in red in Figure 2

$E = E_1 \cup E_2 \cup E_3$ with :

- $E_1 = (A \rightarrow C_1 \cap B \rightarrow C_1) \cap [(AC_1 \cap BC_1) \cup (AC_1 \cap BC_1)]$

- $E_2 = (A \longrightarrow C_1 \cap B \longrightarrow C_2) \cap (AC_1 \cap BC_2)]$
- $E_3 = (A \longrightarrow C_2 \cap B \longrightarrow C_1) \cap (AC_2 \cap BC_1)]$

E_1 corresponds to the case where both agents investigate C_1 but only one of them produces an attack, while E_2 and E_3 are the cases where each agent targets a different argument and only the attack against C_1 succeeds.

Now, let's express the probabilities of E_1 , E_2 and E_3 with and without bias.

First, each event of the form $A \longrightarrow C_i \cap B \longrightarrow C_j$ has a probability of $\frac{1}{4}$. Now, if there is no bias, each agent has the same probability of attacking an argument C_i :

$$\begin{aligned} i \in 1, 2, P(AC_i) &= P(BC_i) = 1 - S_i, \\ i \in 1, 2, P(AC_i) &= P(BC_i) = S_i. \end{aligned}$$

And thus:

$$P(E_1) = \frac{1}{4}[2S_1(1 - S_1)],$$

$$P(E_3) = P(E_2) = \frac{1}{4}(1 - S_1)S_2.$$

Therefore:

$$P(E) = \frac{1}{2}(1 - S_1)(S_1 + S_2).$$

Consider Com' with $bias > 0$. The event E behaves exactly in the same way and it can also be considered as the union of E_1 , E_2 and E_3 . In this case, the probabilities of attacking either argument varies for each agent based on their preference : A is more likely to attack C_2 and less likely to attack C_1 , and vice versa.

$$P_{bias}(AC_1) = 1 - S_1 - bias$$

$$P_{bias}(AC_2) = 1 - S_2 + bias$$

$$P_{bias}(AC_1) = S_1 + bias$$

$$P_{bias}(AC_2) = S_2 - bias$$

$$P_{bias}(BC_1) = 1 - S_1 + bias$$

$$P_{bias}(BC_2) = 1 - S_2 - bias$$

$$P_{bias}(BC_1) = S_1 - bias$$

$$P_{bias}(BC_2) = S_2 + bias$$

Which gives us:

$$\begin{aligned} P_{bias}(E_1) &= \frac{1}{4}[(1 - S_1 - bias)(S_1 - bias) + (1 - S_1 - bias)(S_1 + bias)] \\ &= \frac{1}{2}(S_1 - S_1^2 + bias^2), \end{aligned}$$

$$P_{bias}(E_2) = \frac{1}{4}(1 - S_1 - bias)(S_2 - b),$$

$$P_{bias}(E_3) = \frac{1}{4}(S_2 - bias)(1 - S_1 + bias).$$

By developing and adding each term we obtain

$$P_{bias}(E) = \frac{1}{4}(2S_1 - 2S_1^2 + 2b^2 + S_2 + b - S_1S_2S_1b - S_2b - b^2 + S_2 - S_1S_2 + S_2b - b + bS_1 - b^2),$$

which simplifies to:

$$P_{bias}(E) = \frac{1}{2}(S_1 + S_2 - S_1 - S_2 - S_1S_2) = \frac{1}{2}(1 - S_1)(S_1 + S_2).$$

Notably, $P(E) = P_{bias}(E)$.

Claim 2. Following the same mechanism as before it is possible to show that:

- that $P(Att(C_2) = 0, Att(C_1) = 0) = P_{bias}(Att(C_2) = 0, Att(C_1) = 0)$,
- $P(Att(C_2) = 2, Att(C_1) = 0) = P_{bias}(Att(C_2) = 2, Att(C_1) = 0) + \frac{bias^2}{4}$, and that
- $P(Att(C_2) = 1, Att(C_1) = 1) = P_{bias}(Att(C_2) = 1, Att(C_1) = 1) - \frac{bias^2}{2}$.

Claim 3. Consequently,

$$\mathbb{E}(n_1^1) = 1 \cdot P(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot P(Att(C_2) = 1, Att(C_1) = 1) + 2 \cdot P(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot P(Att(C_2) = 2, Att(C_1) = 0),$$

and

$$\begin{aligned} \mathbb{E}(n_1^{1'}) &= 1 \cdot P_{bias}(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot P_{bias}(Att(C_2) = 1, Att(C_1) = 1) + \\ &+ 2 \cdot P_{bias}(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot P_{bias}(Att(C_2) = 2, Att(C_1) = 0) = \\ &= 1 \cdot P(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot (P(Att(C_2) = 1, Att(C_1) = 1) + \frac{bias^2}{2}) + \\ &+ 2 \cdot P(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot (P(Att(C_2) = 2, Att(C_1) = 0) - \frac{bias^2}{4}) = \\ &1 \cdot P(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot P(Att(C_2) = 1, Att(C_1) = 1) + \\ &+ 2 \cdot P(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot P(Att(C_2) = 2, Att(C_1) = 0) \end{aligned}$$

Thus, $\mathbb{E}(n_1^1) = \mathbb{E}(n_1^{1'})$, i.e. the expected value for the average support is the same for a biased and a non biased community. $\square$

The bias does not induce changes in the agents' expected final distribution. It boils down to showing that although the probabilities of finding attacks are distributed differently, the biases compensate each other and the expected value for n_1^1 is indeed the same in both cases. Thus, we have an indication as to what happens in larger equally distributed communities, that is that when two groups of equal size are biased they cancel any advantage the bias might otherwise grant to one side of the debate.

RÉSUMÉ

Des débats sur les vaccins aux discussions sur le changement climatique, nos sociétés sont régulièrement traversées par des « controverses socio-techniques » mêlant enjeux scientifiques, techniques, juridiques, moraux et politiques. Mon doctorat explore ces débats à travers le prisme de l'argumentation abstraite, une méthode informatique qui simplifie les interactions argumentatives en représentant les arguments comme des entités abstraites. L'objectif principal de ma recherche est d'analyser comment les structures sociales influencent la circulation et l'utilisation des arguments dans les controverses socio-techniques. Cette approche combine des méthodologies théoriques et appliquées, en intégrant aussi bien les dimensions scientifiques que sociopolitiques de ces débats. À partir de simulations multi-agent et d'analyses de controverses réelles, je montre que les controverses sont des processus dynamiques et manipulables, fortement influencés par les structures sociales. Mes travaux mettent en avant la pertinence des outils formels, et en particulier de l'argumentation abstraite, pour l'étude de ces questions complexes.

MOTS CLÉS

Argumentation - Controverses - Modèles multi-agents - Dynamiques des opinions - Epistémologie sociale

ABSTRACT

From debates on vaccines to discussions on climate change, our societies are regularly traversed by “socio-technical controversies” mixing scientific, technical, legal, moral and political issues. My PhD explores these debates through the prism of abstract argumentation, a computational method that simplifies argumentative interactions by representing arguments as abstract entities. The main aim of my research is to analyze how social structures influence the circulation and use of arguments in socio-technical controversies. This approach combines theoretical and applied methodologies, integrating both the scientific and socio-political dimensions of these debates. Using multi-agent simulations and analyses of real-life controversies, I show that controversies are dynamic and manipulable processes, strongly influenced by social structures. My work highlights the relevance of formal tools, and in particular abstract argumentation, for the study of these complex issues.

KEYWORDS

Argumentation - Controversies - Agent-Based Models - Opinion Dynamics - Social Epistemology