JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

An Agent-Based Model of MySide Bias in Scientific Debates

Louise Dupuis de Tarlé[1], Matteo Michelini[2,3], AnneMarie Borg[4], Gabriella Pigozzi[1], Juliette Rouchier[1], Dunja Šešelja[2,3], Christian Straßer[3]

[1] *LAMSADE, Université Paris-Dauphine, 8 rue Winston Churchill, L'Haÿ-les-Roses, 94240, France*
[2] *Ruhr University Bochum, Wasserstr. 221,44799, Bochum, Germany*
[3] *Philosophy & Ethics Group, Technical University of Eindhoven, Atlas 9.328, 5600 MB, Eindhoven, Netherlands*
[4] *Utrecht University, Heidelberglaan 8, 3584 CS, Utrecht, Netherlands*
Correspondence should be addressed to *louise.dupuis@dauphine.eu; matteo.michelini@edu.ruhr-uni-bochum.de*

Journal of Artificial Societies and Social Simulation 27(3) 1, 2024
Doi: 10.18564/jasss.5413 Url: http://jasss.soc.surrey.ac.uk/27/3/1.html

Received: 16-10-2023 Accepted: 11-04-2024 Published: 30-06-2024

Abstract: In this paper, we present an agent-based model for studying the impact of 'myside bias' on the argumentative dynamics in scientific communities. 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. 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 distribution of agents for each alternative at the start of the scientific debate.

Keywords: Myside Bias, Abstract Argumentation, Agent-Based Models, Argumentative Exchange

Introduction

1.1 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 '90s gave rise to various analytical models examining the relationship between individual and group rationality (e.g., Kitcher 1990; Zamora Bonilla 1999; Brock & 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 & Krause 2006; Zollman 2007).

1.2 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).

1.3 This paper aims to examine one such bias which has gained significant attention in cognitive science: what Mercier, Sperber and colleagues define as *myside bias* (Mercier 2017). 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 & 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 favour of held views rather than seeking out defeating arguments. We stick here to the conceptalisation by Mercier, Sperber and colleagues (Mercier & Heintz 2014; Mercier & Sperber 2011; Mercier et al. 2016) that only consider myside bias as a *production* bias, and differentiate from other uses of the notion, concerned either only with the evaluation of arguments or with both production and evaluation (Stanovich & West 2007; Stanovich et al. 2013; Mercier 2017; Wolfe & 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 & Heintz 2014).

1.4 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" (p. 516). Therefore, Mercier & 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.

1.5 Moreover, potential adverse effects of myside bias can be further mitigated when scientists share evaluative standards for what counts as a good argument (Mercier & 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 & Heintz (2014) argue, different scientific domains come with more or less such shared standards. While mathematics comes 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.

1.6 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?

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

1. **Q1**: What kind of effect (if any) does the bias create in a scientific community that exchanges arguments?
2. **Q2**: Do shared beliefs in the form of shared evaluative standards among scientists help to mitigate the detrimental effects of the bias?

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

1.8 To investigate these issues, we develop an ABM[2] based on the abstract argumentation framework (Dung 1995). Abstract argumentation has previously been used for the formal modeling of scientific debates (Šešelja & Straßer 2013) as well as in agent-based modeling of scientific inquiry (Borg et al. 2018, 2019). 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 & Heintz 2014) acts as a mitigating factor against the insidious effects of the bias.

1.9 The paper 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 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 paper.

Myside Bias and Scientific Argumentation

2.1 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.

Myside bias as a production bias

2.2 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.

2.3 **Example 1**. 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 & Weber 2012; Šešelja & 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.

2.4 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 favour 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 2017; Mercier & Sperber 2017). In the drift debate, this may have resulted in 'biased drifters' who were more likely to produce arguments in favour of the explanatory power of Drift than arguments 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).

2.5 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 important 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?

2.6 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 & Straßer 2013) as well as in agent-based modeling of scientific inquiry (Borg et al. 2018, 2019). 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.

2.7 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 & Flache 2013; Borg et al. 2017b; 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 & Hartmann 2022; Gabriel & O'Connor 2024; 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 & Shamon 2023; Singer et al. 2018; Dupuis de Tarlé et al. 2022; Proietti & Chiarella 2023).

2.8 Now, before proceeding to the presentation of our model (in Section 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 1.

Abstract argumentation

2.9 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 (e.g., Gabbriellini & Torroni 2014; Borg et al. 2017a; Dupuis de Tarlé et al. 2022; Butler et al. 2019; Taillandier et al. 2021).

2.10 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 1 (Argumentation framework)**.** An **argumentation framework** (AF) is a pair $\langle \mathscr{A}, \mathscr{R} \rangle$ where $\mathscr{A}$ is a finite and non-empty set of (abstract) arguments, and $\mathscr{R} \subseteq \mathscr{A} \times \mathscr{A}$ is a binary relation on $\mathscr{A}$ called the attack relation. For two arguments $a, b \in \mathscr{A}$, we say that a attacks b in case $(a, b) \in \mathscr{R}$.

Definition 2 (Path)**.** Let $\langle \mathscr{A}, \mathscr{R} \rangle$ be an AF and $a_1, a_2, ..., a_n \in \mathscr{A}$ be n arguments, with $n \in \mathbb{N}$. We say that $(a_1, a_2, ..., a_n)$ is a **path** from a_1 to a_n if for every $i \in [\![1; n]\!]$, $(a_i, a_{i+1}) \in \mathscr{R}$.

Definition 3 (Attacker and Defender)**.** Let $\langle \mathscr{A}, \mathscr{R} \rangle$ be an AF and $a, b \in \mathscr{A}$ be two arguments. Argument 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.

2.11 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. In this paper we use the grounded set due to its uniqueness. [3]

Definition 4 (Grounded set)**.** Let AF $= \langle \mathscr{A}, \mathscr{R} \rangle$ be an AF. An argument $a \in \mathscr{A}$ is **acceptable with respect to a set of arguments** E in case for every b for which $(b, a) \in \mathscr{R}$, there is a $c \in E$ for which $(c, b) \in \mathscr{R}$. 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 $\mathscr{A}$ 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.

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

2.13 **Example 2**. Imagine how a stylized dicussion between Alfred Wegener, proponent of 'Drift', and Ernest Ingersoll, proponent of 'Permanentism', may have played out (see Example 1). 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 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.

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

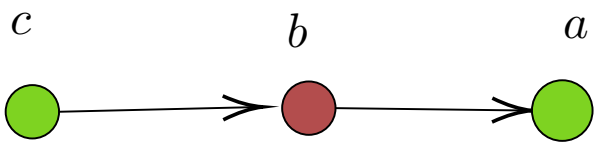

Figure 1: Representation of the debate of Example 2 as an argumentation graph, where the arguments are represented as nodes and the attack relations as directed edges. Arguments a, c in light green are acceptable, and b in dark red is not.

The Model

3.1 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.

Modelling opposing views with argumentation frameworks

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

Definition 5 (Research Programme)**.** A **research programme (**$ResProg$**)** is a set of arguments of size N_{CA} (≥ 1).

3.3 We call the arguments included in a research programme *central arguments* (hence, N_{CA} is the number of central arguments per program). Central arguments provide the central theoretical tenets of a research programme. They give the argumentative support to the core hypothesis of the program. For example, in the continental drift research programme 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 programmes contain the same number of distinct central arguments.

3.4 Agents engage in a debate by discussing the central arguments of two opposing research programmes, $ResProg_1$ and $ResProg_2$. For this, they produce additional (dialetical) arguments that are not contained in research programmes and which are used to attack and/or defend central arguments. For example, the argument concerned with the similarity of flora and fauna can be attacked by claiming that such similarity is the result of the adaptation to similar climate and geological circumstances, and not of the fact that continents were once united.

3.5 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 6 (View of Debate)**.** The view of the debate V_i^t of agent i at step t is an argumentation framework $\langle \mathscr{A}_i^t, \mathscr{R}_i^t \rangle$ such that for any t, i: $ResProg_1, ResProg_2 \subset \mathscr{A}_i^t$.

3.6 As such, the view of an agent always contains the central arguments of each research programme, 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 programme is the root of an in-tree, that is a tree where all edges point towards the root. Second, each argument $a \in A_i$ 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 2 presents an example of an in-tree with its own root, and Figure 3 presents an example of a view. While the view of every agent contains the central arguments of each research programme, different agents can have different views.

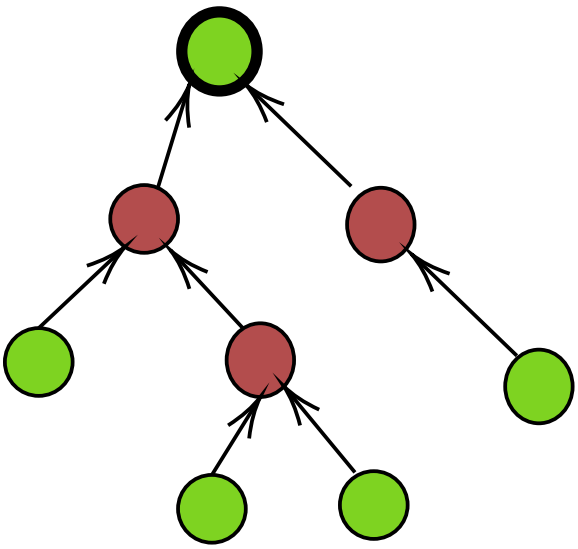

Figure 2: Example of an in-tree. The acceptable arguments are colored in light green, the others in dark red. The central argument is outlined in bold.

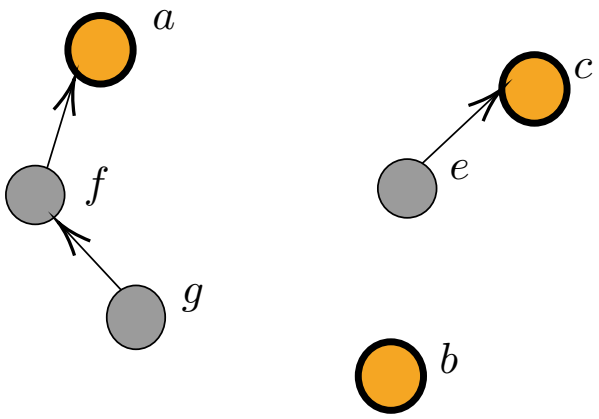

Figure 3: Schematic representation of a possible view of an agent composed of three in-trees with central arguments a, b and c (in orange and bold).

3.7 Based on her own view of the debate, an agent assigns a score to each research programme, 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.

Definition 7 (Score of a Research Programme)**.** 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

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

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

3.9 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 the strength of a, that is $S(a)$. 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 programme have the same strength. This modeling choice allows us to decide which (if any) of the two research programmes is objectively stronger and, consequently, objectively preferable. A more robust research programme represents a research programme that is based on stronger evidence or provides better explanations than the rival one.

3.10 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 programmes S_1, S_2 are treated as simulation parameters.[5]

3.11 Our model simulates an ongoing debate among scientists about the merits of two research programmes $ResProg_1$ and $ResProg_2$. It does so through a succession of T steps in which agents generate new argumentative attacks and update their views and preferences for research programmes. 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.

Baseline protocol

3.12 **Initialization**. Each agent i is initialized with a view whose argument set contains only the central arguments of both research programmes and no attack relation, i.e., $V_i^0 = \langle \mathscr{A}_i^0, \mathscr{R}_i^0 \rangle$ such that $\mathscr{A}_i^0 = ResProg_1 \cup ResProg_2$ and $\mathscr{A}_i^0 = \emptyset$ (see Figure 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 $ResProg_2$.

3.13 **Protocol**. Each step has three phases.

3.14 *Phase 1: "Attack generation".*

- Each agent i randomly chooses some argument a among those present in her set of arguments $\mathscr{A}_i^t$.
- 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]

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 3.21).

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 her view. Every agent j who does not know argument a, that is $a \notin \mathscr{A}_j$, 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 programme based on her view (Definition 7) 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.

3.15 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 programme, and a new argument can only attack an argument that is already present. As a consequence, each central argument of a research programme 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 Borg et al. 2018; Proietti & 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.

3.16 Our agents evaluate the debate impartially, insofar as they always support the research programme with the highest score. As mentioned, 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.

3.17 **Example 3**. Let us illustrate this protocol with the first step of a debate concerning two research programmes of strength $S_1 = 0.5$ and $S_2 = 0.8$. Each research programme contains two central arguments, as shown in Figure 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 $t = 0$ (V_1^0) is shown in Figure 4.

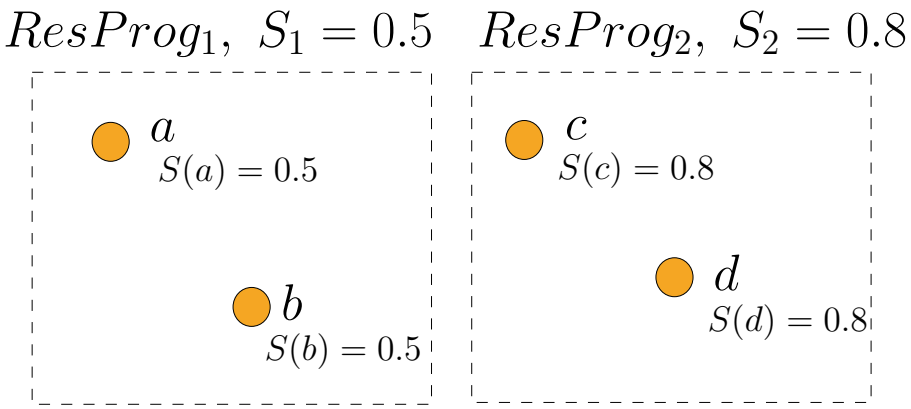

Figure 4: View V_1^0 of the debate for agent 1 at the start of the simulation.

3.18 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 $t = 1$ (V_1^1) shown in Figure 5 is the result of updating the initial view by adding the new argument e and an attack between e and a.

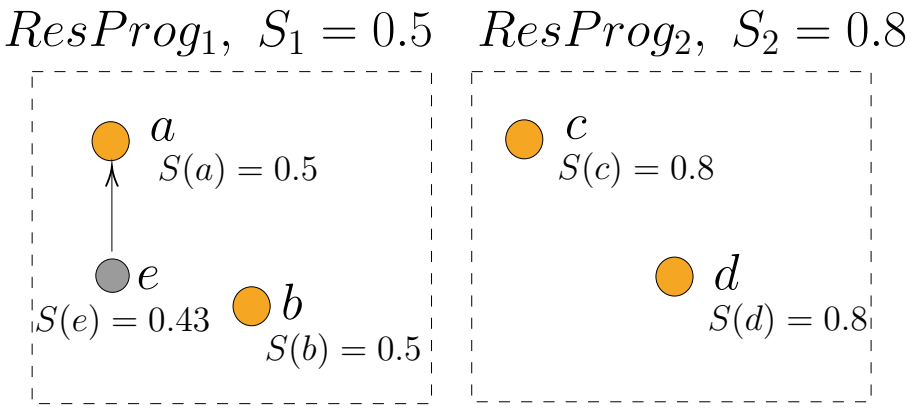

Figure 5: View V_1^1 of the debate.

3.19 Then, $Agent_1$ computes the score of each research programme 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 programme 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.

MySide bias in the production of arguments

3.20 As mentioned above, myside bias is modelled as a factor that makes it easier to *produce* arguments in favour of one's standpoint. In our protocol, the bias increases the probability of generating a counterargument in case the counterargument will *favour* the agent's preferred research programme. An argument favours one's preferred research programme if it either defends the program or attacks the opposing one (recall Definition 3).

Definition 8 (Against - In favour of)**.** An argument is said to be **in favour** of a research programme $ResProg_i$ 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 $ResProg_j$.

3.21 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 favour* of her preferred research programme, then

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

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

$$p_{attack} = \begin{cases} 1 - S(a) + bias & \text{if } 1 - S(a) + bias \leq 1, \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

3.22 The bias parameter changes the probability with which an agent finds a counterargument 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.

3.23 **Example 4**. To illustrate the difference induced by the bias, consider the first step of the same debate of Example 3, where $Agent_1$ is now biased with $bias = 0.3$. The view of $Agent_1$ is the same at the beginning (Figure 4), and she begins the simulation with a preference for $ResProg_1$. $Agent_1$ randomly selects an argument from V_1^0, in this case a. Attacking a would be *against* the agent's preferred research programme, $ResProg_1$. 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. $Agent_1$ computes the score of each research programme 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 programmes for $Agent_1$ are the same: $score(ResProg_1) = 2$ and $score(ResProg_2) = 2$. Since $score(ResProg_1) = score(ResProg_2)$, agent 1 keeps preferring $ResProg_1$ in the next round.

3.24 This example highlights how the bias makes it harder to generate arguments against one's preferred research programme, 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 programme. In this case, it is as though her bias prevents $Agent_1$ from seeing the faults of her preferred point of view.

3.25 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 & 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 programmes.

Shared beliefs as shared standards of evaluation

3.26 Mercier & 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.

3.27 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]

3.28 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 her view.

3.29 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.

3.30 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.

Simulations and observation

4.1 In our simulation, a community of agents investigates two research programmes by producing new arguments. Each research programme 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.

4.2 Each scientist always supports the research programme that appears stronger to her, i.e., the one that has the highest score in her view. 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.

4.3 Now, we shall summarize the parameters of our simulation (see Table 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 engages 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 programme (N_{CA}) and the coefficient that determines the standard deviation of the distribution from which the strength of a new argument is drawn ($StDev$). It is given by a pair $Env = \langle N_{CA}, StDev \rangle$. Finally, there are the two research programmes, $ResProg_1$ and $ResProg_2$, that come with their respective strengths S_1, S_2.

Table 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 programme.	$[1, 10]$	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 $ResProg_2$.	$[0, 1]$	-

Epistemic evaluation

4.4 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 programme for different values of the strengths of the research programmes (S_1, S_2).

4.5 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 programme $ResProg_i$ 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 programme, 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 programme'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 programmes ($d = S_1 - S_2$ 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 or equal to $ResProg_2$, and negative otherwise.

4.6 A very clear indicator of the epistemic success of a community would be if every member always preferred the strongest research programme. 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.

4.7 We report the *average* support for a research programme 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 programmes 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 programme at 100 steps (i.e., at $t = 100$), which is a reasonable convergence time, as explained in Section 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 9 (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 programme is higher than half the size of the community. We say that Com is epistemically successful for an interval $D = [d_1, d_2]$ and for an environment Env iff it is epistemically successful for all $d \in D$ and for Env.

4.8 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.

4.9 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 1 such community is epistemically successful.

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

$$PolRat^t = \frac{\min(n_1^t, n_2^t)}{\max(n_1^t, n_2^t)} \tag{3}$$

4.11 Following Bramson et al. (2017)'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 5.11. 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.

Results

5.1 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.

5.2 Unless specified otherwise, all the following plots and observations are made using the default values for the parameters in Table 1.

5.3 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.

Time convergence

5.4 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 programme, 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 agents generate a set of arguments that leads one or more agents to change their preference(s).

5.5 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 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]

5.6 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.

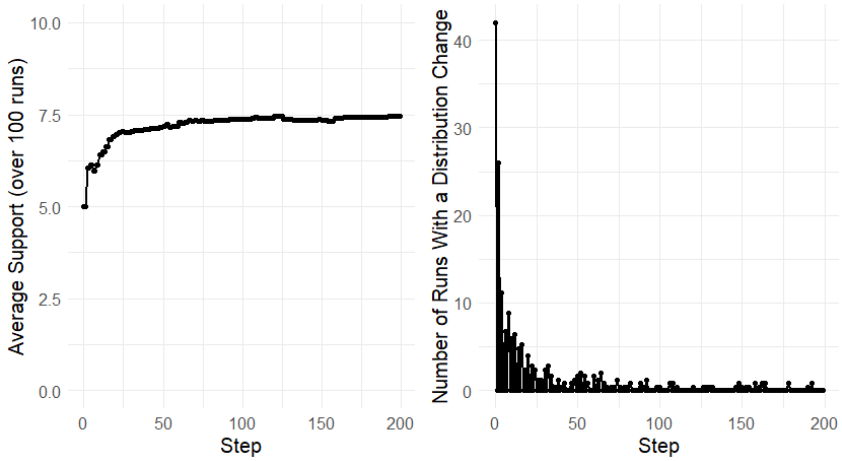

Figure 6: On the left: the evolution of the value for the average support for $ResProg_1$ is plotted at any step. On the right: the number of runs (y-axis) in which the value of the support changes in the time-step represented on the x-axis. The total number of runs is one hundred in both cases. Here, $d = 0.1, n_1^0 = 5$. Similar plots can be obtained for the whole parameter space.

The baseline model

5.7 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 programme. 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 programmes ($d = S_1 - S_2$) increases.

5.8 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, 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.

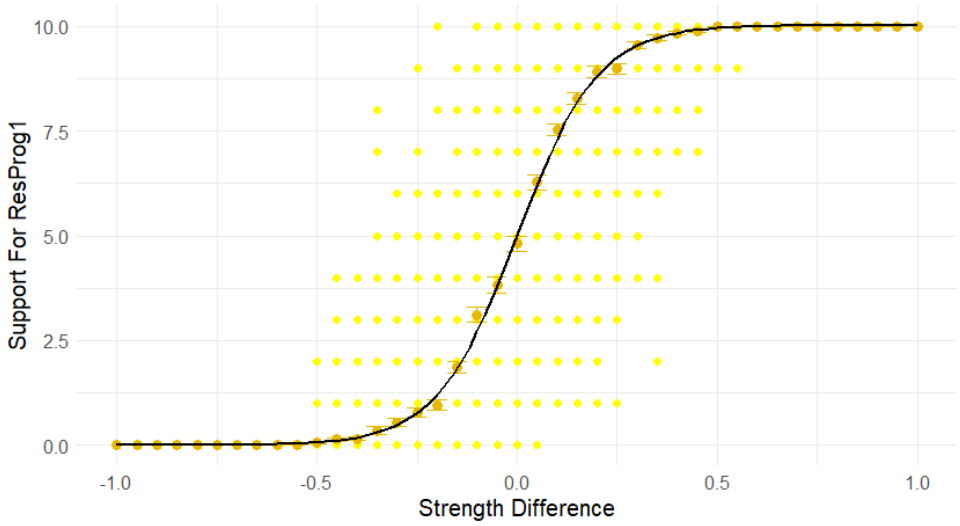

Figure 7: The support for $ResProg_1$ is plotted against the strength difference d. The light small yellow points represent the value of support $Supp$ for singular runs. The ocher ones are the values for the average support ($\overline{Supp}_d$). The standard error is represented as an interval around $\overline{Supp}_d$. The black line is the graph of the function $y = {}^{10}/_{(1+e^{-9.3d})}$, which is the best approximation of our distribution.

5.9 The individual data points may be found quite far away from the averaged values (Figure 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.

5.10 The relationship between the average support ($\overline{Supp}_d$) and the strength difference (d) shown in Figure 7 can be approximated by a *sigmoid* model with formula:

$$\overline{Supp}_d = \frac{\alpha}{1 + e^{-\beta \cdot d}} \tag{4}$$

with $\alpha = 10, \beta \in [5, 11]$.[9] The value α is the maximum of the curve: it is equal to the total number of agents, whereas β controls the steepness of the curve and depends on the value of the parameters (Table 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.

5.11 The emergence of the sigmoid is robust for any parameter combination in Table 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).

5.12 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 programmes. If an agent only includes a part of all the produced arguments in its view, that part is more likely to be misleading.

5.13 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.

A biased community

5.14 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.

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.

The detrimental effect of the bias

5.15 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 programmes, 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 ($n_1^0 > N/2$) the average support for $ResProg_1$ will be higher than if the community was not biased, while if the initial support for $ResProg_1$ is a minority ($n_1^0 < N/2$), the final average support for $ResProg_1$ will be lower (see Figure 8). Therefore, if there is a clear majority preferring one side at the beginning of the debate ($n_1^0 \neq N/2$), a biased community is not fully epistemically successful: the average support will be higher for the weaker research programme for a certain interval of values of the strength difference.[10] For example, if nine out of ten agents start our preferring $ResProg_1$ (as in the green line in Figure 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 $d = ResProg_1 - ResProg_2 = -0.2$).

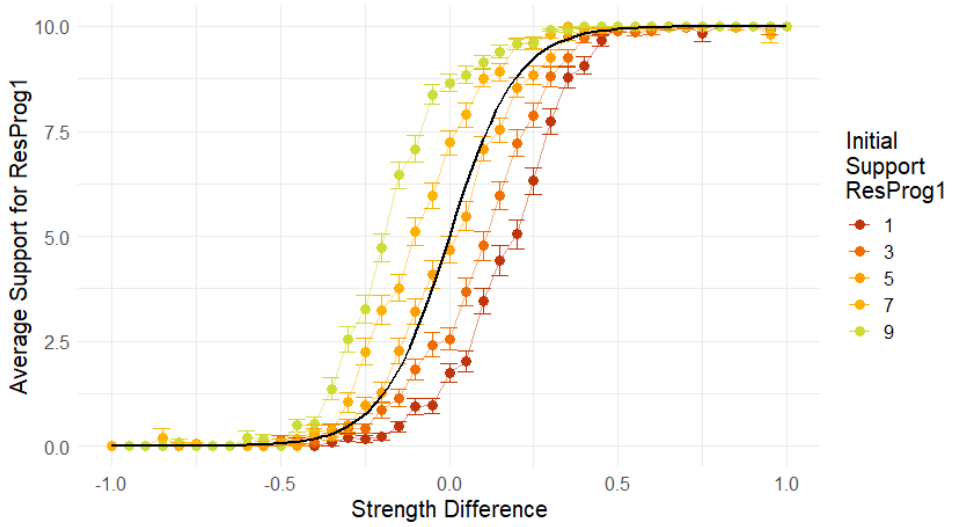

Figure 8: The average support for $ResProg_1$ is plotted against the difference d. The agents all have a bias of 0.3. The colours represent initial support for $ResProg_1$. The standard error is represented as an interval around $\overline{Supp_d}$. The black line represents the average support produced by the baseline model (Equation 6).

5.16 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 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 10).[11]

5.17 The results of a biased community can also be approximated by Equation 6. To do so, we introduce a generalized notion of strength difference, which allows us to use Equation 6 to fit the results of biased and unbiased communities. We call this refined notion of strength difference *biased strength difference*.

Definition 10 (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 : $d_{bias} = {}^{n_1^0}/_N \cdot (\min(S_1 + bias, 1) - \max(S_2 - bias, 0)) + {}^{(N-n_1^0)}/_N \cdot (\max(S_1 - bias, 0) - \min(S_2 + bias, 1))$.

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

5.19 **Observation 1**. Using the biased strength difference (d_{bias}), we can describe all the results obtained so far (that is, those in Figure 8 and Figure 7) as follows:

$$\overline{Supp}_{d_{bias}} = \frac{\alpha}{1 + e^{-\beta \cdot d_{bias}}}, \tag{5}$$

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

5.20 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 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 programmes 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 programmes 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 programmes as having different strengths, which are influenced not only by the bias but also by the number of people supporting each theory initially.

Equally distributed biased communities

5.21 Our results show that the community Com is fully epistemically successful when there is an equal number of agents preferring each research programme at the beginning ($n_1^0 = n/2$). We call these communities "equally distributed", referring to their *initial* distribution of support. Consider again Figure 8: when the initial support for $ResProg_1$ 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.

5.22 These results align with the conjecture by Mercier & 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.

5.23 In summary, biased communities are not epistemically successful, as the bias makes communities more inclined to produce a higher average support for the research programme 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]

More bias, less polarization

5.24 Finally, we also discuss the effect of bias on the polarization of a community. As can be noted from Figure 9 more biased communities tend to polarize less, that is to exhibit a lower average polarization ratio.

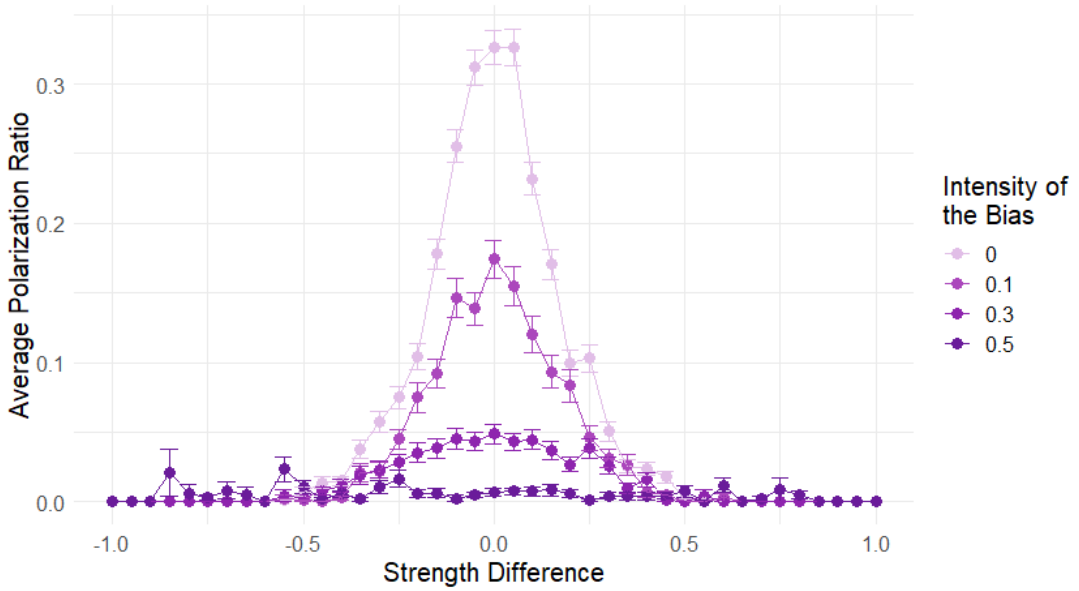

Figure 9: The average polarization ratio plotted against the strength difference d. The colours represent different values for the bias.

5.25 Although this result may perhaps seem 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 programme 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.

Results with shared beliefs

5.26 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 behaviour of the biased community to that of the baseline community, the more the shared belief filter has a mitigating effect.

5.27 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 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. 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).

5.28 In addition, when the bias is quite low, the effect of the shared beliefs is almost invisible: consider the first line of Figure 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.

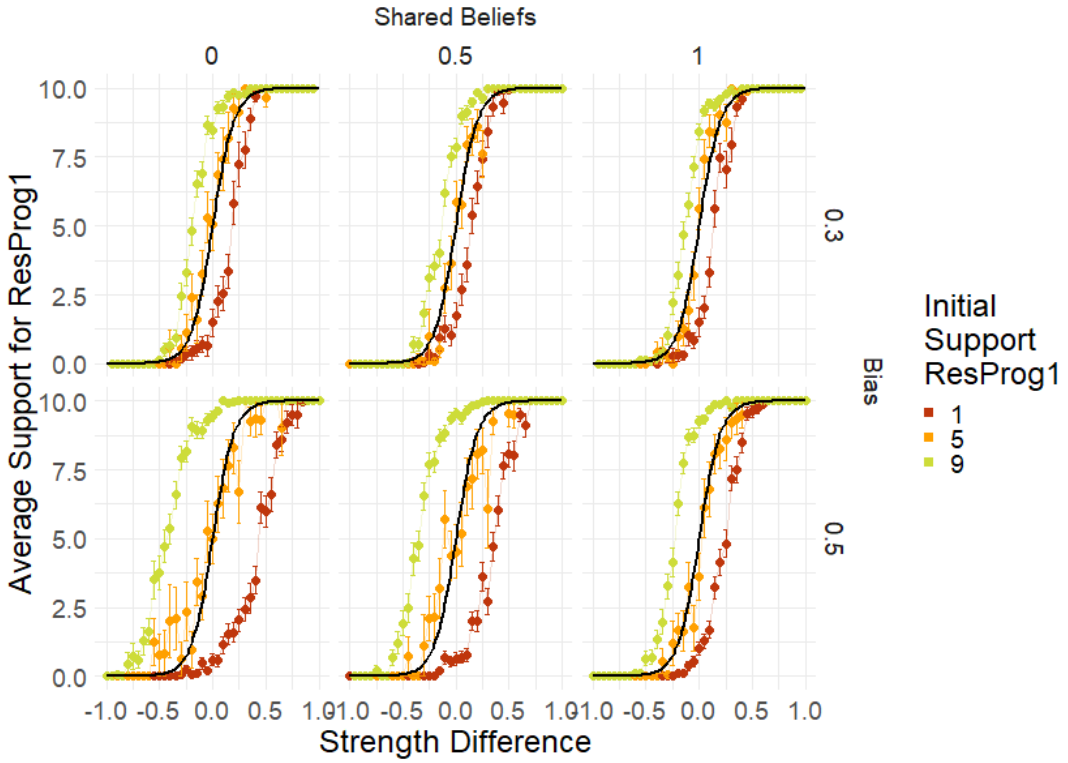

Figure 10: The average support for $ResProg_1$ plotted against d. The colours represent the initial support for $ResProg_1$. The standard error is represented as an interval around the $\overline{Supp}$. The subplots corresponds to different intensity of shared beliefs and of the bias. The black line represents the average support produced by the baseline model (Equation 6).

5.29 Similarly, shared beliefs are ineffective in mitigating the bias's impact when the difference between the strength of the two research programmes is very small (d close to 0). This can be seen in Figure 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.

Related Works

6.1 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.

6.2 Previous ABMs based on abstract argumentation have been used to study opinion dynamics and polarization effects in argumentative settings (Gabbriellini & Torroni 2014; Banisch & Olbrich 2021; Butler et al. 2019; Taillandier et al. 2021; Mäs & Flache 2013; Proietti & Chiarella 2023; Singer et al. 2018), gradual learning in debates (Dupuis de Tarlé et al. 2022), 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.

6.3 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 (Banisch & Olbrich 2021; Mäs & Flache 2013; Taillandier et al. 2021; Dupuis de Tarlé et al. 2022; Proietti & 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).

6.4 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.

6.5 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 & Zhang (2016), Banisch & Shamon (2023), Proietti & Chiarella (2023), Dupuis de Tarlé et al. (2022) 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 & O'Connor (2024), Baccini & 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 & Chiarella (2023) and Singer et al. (2018) biased agents tend to forget arguments opposed to their present views sooner than the other ones.

6.6 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. 2017), and a specific setting in which the bias can emerge.

6.7 Among the many articles that study the effect of biases in deliberative settings, Gabriel & O'Connor (2024), 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 & 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 & O'Connor (2024) 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.

Conclusion and Outlook

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

7.2 Following Mercier & Heintz (2014) and Mercier (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 programme if such a research programme 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 programme 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 & Flache (2013), Proietti & Chiarella (2023), Singer et al. (2018) who find that biased communities polarize more.

7.3 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 programmes have a similar strength. In this way we confirmed and qualified a conjecture by Mercier & Heintz (2014).

7.4 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 & 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.

7.5 Our findings hinge on a number of assumptions. First, we assume the objective strength of the research programmes 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 programme 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.

7.6 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 & 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.

7.7 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.

7.8 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 & 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 & Tetlock (2021).

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

7.10 As noted by Mercier & Heintz (2014), 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.

Appendix A: 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 unnacceptable.

In our setting, the arguments which are not attacked are the leafs of each argumentation tree.

Algorithm 1 Algorithm to compute the acceptable arguments of C

C is an argumentation tree.

```
G ← leaf(C)
O ← empty list
while length of O + length of G < length of C do
  for all n in G do
    O ← O+ successors of n
  end for
  for all o in O do
    for a in successors of o do
      if a not in O and predecessors of a is included in O then
        G ← G + {a}
      end if
    end for
  end for
end while
```

Appendix B: Additional Results

Table 2 register how different parameters of $DebSet$ affect the coefficient b of the function

$$\overline{Supp} = \frac{\alpha}{1 + e^{-\beta \cdot d}}. \tag{6}$$

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 2: 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.

Appendix C: Analytical Analysis of Equally Distributed Communities

To substantiate the (quite surprising) computational finding that biased equally distributed communities are fully epistemically successful, we provide an analytical result about such communities. 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 expected number of agents supporting $ResProg_1$ in a case in which agents are biased is the same as in a case in which agents are not 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 $N = N', n_1^0 = n_1^{0\prime}, p_{see} = p'_{see} = 1$ 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 $\mathbb{E}(n_1^{1\prime})$. 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) $P(Att(C_2) = 1, Att(C_1) = 1) = P_{bias}(Att(C_2) = 1, Att(C_1) = 1) - \frac{bias^2}{2}$.
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 biased agents and unbiased ones. Figure 11 presents a schematic view of the probability tree associated with this problem. We use the following names of events.

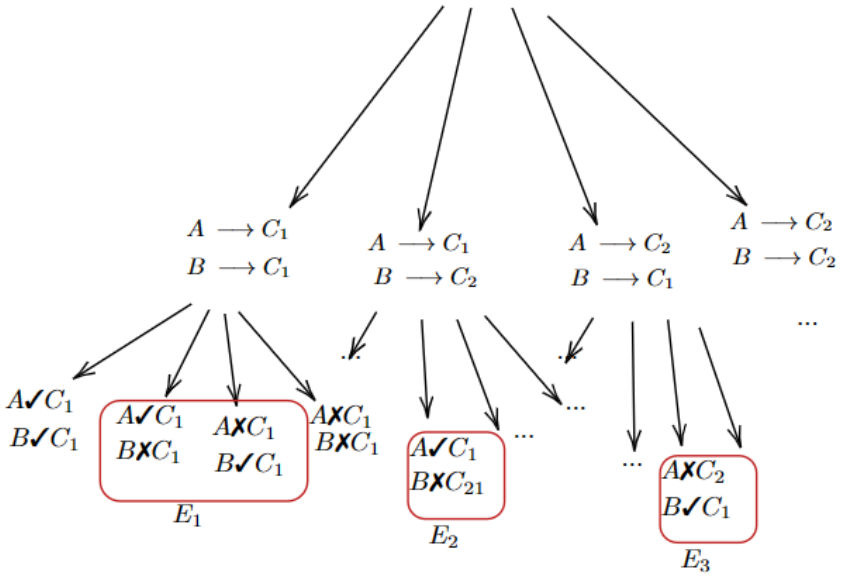

Figure 11: Representation of the probability tree.

- A (resp. B) $\longrightarrow C_i$: Agent A (resp. B) is investigating argument C_i.
- A (resp. B) $\checkmark C_i$: Agent A (resp. B) has produced an attack against C_i.
- A (resp. B) $\cancel{\checkmark} C_i$: 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 11

$E = E_1 \cup E_2 \cup E_3$ with :

- $E_1 = (A \longrightarrow C_1 \cap B \longrightarrow C_1) \cap [(A\checkmark C_1 \cap B\cancel{\checkmark} C_1) \cup (A\cancel{\checkmark} C_1 \cap B\checkmark C_1)]$
- $E_2 = (A \longrightarrow C_1 \cap B \longrightarrow C_2) \cap (A\checkmark C_1 \cap B\cancel{\checkmark} C_2)]$
- $E_3 = (A \longrightarrow C_2 \cap B \longrightarrow C_1) \cap (A\cancel{\checkmark} C_2 \cap B\checkmark C_1)]$

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:

$i \in 1,2, P(A\checkmark C_i) = P(B\checkmark C_i) = 1 - S_i,$
$i \in 1,2, P(A\cancel{\checkmark} C_i) = P(B\cancel{\checkmark} C_i) = S_i.$

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 her preference: A is more likely to attack C_2 and less likely to attack C_1, and vice versa.

$P_{bias}(A\checkmark C_1) = 1 - S_1 - bias$
$P_{bias}(A\checkmark C_2) = 1 - S_2 + bias$
$P_{bias}(A\text{✗}C_1) = S_1 + bias$
$P_{bias}(A\text{✗}C_2) = S_2 - bias$

$P_{bias}(B\checkmark C_1) = 1 - S_1 + bias$
$P_{bias}(B\checkmark C_2) = 1 - S_2 - bias$
$P_{bias}(B\text{✗}C_1) = S_1 - bias$
$P_{bias}(B\text{✗}C_2) = S_2 + bias$

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

$$P_{bias}(E) = \frac{1}{4}(2S_1 - 2S_1^2 + 2b^2 + S_2 + b - S_1S_{2S1}b - S_2b - b^2 + S_2 - S_1S_2 + S_2b - b + bS_1 - b^2),$$

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

$$\begin{aligned}\mathbb{E}(n_1^{1\prime}) &= 1 \cdot P_{bias}(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot P_{bias}(Att(C_2) = 1, Att(C_1) = 1) + \\ &+ 2 \cdot P_{bias}(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot P_{bias}(Att(C_2) = 2, Att(C_1) = 0) = \\ &= 1 \cdot P(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot (P(Att(C_2) = 1, Att(C_1) = 1) + \frac{bias^2}{2}) + \\ &+ 2 \cdot P(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot (P(Att(C_2) = 2, Att(C_1) = 0) - \frac{bias^2}{4}) = \\ &1 \cdot P(Att(C_2) = 0, Att(C_1) = 0) + 1 \cdot P(Att(C_2) = 1, Att(C_1) = 1) + \\ &+ 2 \cdot P(Att(C_2) = 1, Att(C_1) = 0) + 2 \cdot P(Att(C_2) = 2, Att(C_1) = 0)\end{aligned}$$

Thus, $\mathbb{E}(n_1^1) = \mathbb{E}(n_1^{1\prime})$, i.e., the expected value for the average support is the same for a biased and a non biased community.

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.

Notes

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

[2]A 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/`

[3]It 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 Section 3.1).

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

[5]Note that the strength of each research programme is not accessible to the agents and thus does not directly impact their preferences. However, it affects agents' preferences indirectly via the influence it has on the process of argument generation.

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

[7]For 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.

[8]We 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.

[9]For the default values, $\beta = 9.3$ and the residual standard error is 0.1045 on 39 degrees of freedom, i.e., 39 different values for d: $0, 0.05, 0.1, 0.15, ...1$.

[10]We 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 $bias = 0$.

[11]Although this behaviour can be observed through the whole parameter space, i.e., for any possible parameter combination studied (see Table 1), it is worth highlighting the slightly different behaviour of what we call *Limit cases.* As the probability of attacking a research programme 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 programme 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 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 $S_2 = 0.35$.

[12]Again, 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 $d_{bias}: 0, 0.05, 0.1, 0.15, ...1$.

[13]This is the case under the assumption that we do not incur in 'limit cases' - see Endnote 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.

References

Baccini, E., Christoff, Z., Hartmann, S. & Verbrugge, R. (2023). The wisdom of the small crowd: Myside bias and group discussion. *Journal of Artificial Societies and Social Simulation*, *26*(4), 7

Baccini, E. & Hartmann, S. (2022). The myside bias in argument evaluation: A Bayesian model. Proceedings of the Annual Meeting of the Cognitive Science Society

Banisch, S. & Olbrich, E. (2021). An argument communication model of polarization and ideological alignment. *Journal of Artificial Societies and Social Simulation*, *24*(1)

Banisch, S. & Shamon, H. (2023). Biased processing and opinion polarization: Experimental refinement of argument communication theory in the context of the energy debate. *Sociological Methods & Research*, *2023*

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, *1*(3), 221–235

Borg, A., Frey, D., Šešelja, D. & Straßer, C. (2017a). An argumentative agent-based model of scientific inquiry. In S. Benferhat, K. Tabia & M. Ali (Eds.), *Advances in Artificial Intelligence: From Theory to Practice*, (pp. 507–510). Cham: Springer International Publishing

Borg, A., Frey, D., Šešelja, D. & Straßer, C. (2017b). Examining network effects in an argumentative agent-based model of scientific inquiry. In A. Baltag, J. Seligman & T. Yamada (Eds.), *Logic, Rationality, and Interaction*, (pp. 391–406). Berlin Heidelberg: Springer

Borg, A., Frey, D., Šešelja, D. & Straßer, C. (2019). Theory-choice, transient diversity and the efficiency of scientific inquiry. *European Journal for Philosophy of Science*, *9*(2), 1–25

Borg, A. M., Frey, D., Šešelja, D. & Straßer, C. (2018). Epistemic effects of scientific interaction: Approaching the question with an argumentative agent-based model. *Historical Social Research*, *43*(1), 285–309

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

Brock, W. A. & Durlauf, S. N. (1999). A formal model of theory choice in science. *Economic Theory*, *14*, 113–130

Butler, G., Pigozzi, G. & Rouchier, J. (2019). An opinion diffusion model with deliberation. 20th International Workshop on Multi-Agent-Based Simulation (MABS 2019)

Clark, C. & Tetlock, P. (2021). Adversarial collaboration: The next science reform. In C. L. Frisby, R. E. Redding, W. T. O'Donohue & S. O. Lilienfeld (Eds.), *Ideological and Political Bias in Psychology*, (pp. 905–927). Berlin Heidelberg: Springer

Douglas, H. (2009). *Science, Policy, and The Value-Free Ideal*. Pittsburgh, PA: 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., Bonzon, E. & Maudet, N. (2022). Multiagent dynamics of gradual argumentation semantics. Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems

Dutilh Novaes, C. (2018). The enduring enigma of reason. *Mind & Language*, *33*(5), 513–524

Fu, G. & Zhang, W. (2016). Opinion formation and bi-polarization with biased assimilation and homophily. *Physica A: Statistical Mechanics and its Applications*, *444*, 700–712

Gabbriellini, S. & Torroni, P. (2014). A new framework for ABMs based on argumentative reasoning. In B. Kamiński & G. Koloch (Eds.), *Advances in Social Simulation*, (pp. 25–36). Berlin Heidelberg: Springer

Gabriel, N. & O'Connor, C. (2024). Can confirmation bias improve group learning? Available at: `http://philsci-archive.pitt.edu/20528/`

Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research Online*, *2*(2), 1–15

Hegselmann, R. & Krause, U. (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

Kitcher, P. (1990). The division of cognitive labour. *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), 4

Kuhn, T. (2000). *The Road Since Structure*. Chicago, IL: The University of Chicago Press

Liu, Q., Zhao, J. & 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, NJ: Princeton University Press

Mäs, M. & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLoS One*, *8*(11), e74516

Mercier, H. (2017). Confirmation bias - Myside bias. In R. F. Pohl (Ed.), *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment and Memory*, (pp. 109–124). London: Routledge

Mercier, H., Bonnier, P. & Trouche, E. (2016). Why don't people produce better arguments? In L. Macchi, M. Bagassi & R. Viale (Eds.), *Cognitive Unconscious and Human Rationality*, (p. 205). Cambridge, MA: The MIT Press

Mercier, H. & Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi*, *33*(2), 513–524

Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74

Mercier, H. & Sperber, D. (2017). *The Enigma of Reason*. Cambridge, MA: Harvard University Press

Mutz, R., Bornmann, L. & 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), e48509

Pera, M. (1994). *The Discourses of Science*. Chicago, IL: The University of Chicago Press

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

Peters, U. (2020). Illegitimate values, confirmation bias, and mandevillian cognition in science. *The British Journal for the Philosophy of Science*, *72*(4)

Popper, K. R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Londo: Routledge

Proietti, C. & 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

Rescher, N. (2007). *Dialectics*. Berlin: De Gruyter

Šešelja, D. & Straßer, C. (2013). Abstract argumentation and explanation applied to scientific debates. *Synthese*, *190*, 2195–2217

Šešelja, D. & 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

Singer, D. J., Bramson, A. L., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A. & Berger, W. J. (2018). Rational social and political polarization. *Philosophical Studies*, *176*, 2243–2267

Smart, P. R. (2018). Mandevillian intelligence. *Synthese*, *195*(9), 4169–4200

Stanovich, K. E. & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247

Stanovich, K. E., West, R. F. & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264

Taillandier, P., Salliou, N. & 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

Wolfe, C. R. & Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & Reasoning*, *14*(1), 1–27

Zamora Bonilla, J. (1999). The elementary economics of scientific consensus. *Theoria: An International Journal for Theory, History and Foundations of Science*, (pp. 461–488)

Zollman, K. (2011). Computer simulation and emergent reliability in science. *Journal of Artificial Societies and Social Simulation*, *14*(4), 15

Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587