PAPER IN FORMAL METHODS AND EXACT SCIENCES

Theory-choice, transient diversity and the efficiency of scientific inquiry

AnneMarie Borg[1] · Daniel Frey[2] · Dunja Šešelja[3] (iD) · Christian Straßer[1] (iD)

Abstract

Recent studies of scientific interaction based on agent-based models (ABMs) suggest that a crucial factor conducive to efficient inquiry is what Zollman (2010) has dubbed 'transient diversity'. It signifies a process in which a community engages in parallel exploration of rivaling theories lasting sufficiently long for the community to identify the best theory and to converge on it. But what exactly generates transient diversity? And is transient diversity a decisive factor when it comes to the efficiency of inquiry? In this paper we examine the impact of different conditions on the efficiency of inquiry, as well as the relation between diversity and efficiency. This includes certain diversity-generating mechanisms previously proposed in the literature (such as different social networks and cautious decision-making), as well as some factors that have so far been neglected (such as evaluations underlying theory-choice performed by scientists). This study is obtained via an argumentation-based ABM (Borg et al. 2017, 2018). Our results suggest that cautious decision-making does not always have a significant impact on the efficiency of inquiry while different evaluations underlying theory-choice and different social networks do. Moreover, we find a correlation between diversity and a successful performance of agents only under specific conditions, which indicates that transient diversity is sometimes not the primary factor responsible for efficiency. Altogether, when comparing our results to those obtained by structurally different ABMs based on Zollman's work, the impact of specific factors on efficiency of inquiry, as well as the role of transient diversity in achieving efficiency, appear to be highly dependent on the underlying model.

This article is part of the Topical Collection on EPSA17: Selected papers from the biannual conference in Exeter
Guest Editors: Thomas Reydon, David Teira, Adam Toon

✉ Dunja Šešelja
   Dunja.Seselja@lrz.uni-muenchen.de

Extended author information available on the last page of the article

1 Introduction

Recent studies on epistemic effects of scientific interaction, conducted via agent-based models (ABMs), have largely focused on the context of theoretical diversity, where a scientific community pursues different rivaling theories within a given scientific domain (Borg et al. 2017, 2018; Frey and Šešelja 2018a; Grim 2009; Grim et al. 2013; Kummerfeld and Zollman 2016; Zollman 2007, 2010). Since one of the rivaling theories is assumed to be the best, agents are successful if they manage to converge on it. A take-home message from a number of these studies has been the following: in order for an inquiry to be successful it needs a property of 'transient diversity' (Zollman 2010). Transient diversity refers to a process in which a community engages in a parallel exploration of different theories, which lasts sufficiently long to prevent a premature abandonment of the best of the available theories, but which eventually gets replaced by a consensus on the best theory. Or as Pöyhönen and Kuorikoski (2016) specify it, transient diversity represents "a proper balance between the diversity of beliefs and consensus".

But what does exactly generate this kind of balance? Zollman has suggested that transient diversity can be obtained either by limiting information flow among scientists or by equipping them with extreme prior values for their initial hypotheses (though not by both of these mechanisms at the same time). Kummerfeld and Zollman (2016) suggest that institutional encouragement of unpopular, risky paths of inquiry may be necessary to obtain such a diversity. Finally, Frey and Šešelja (2018a) suggest that cautious decision-making may be yet another mechanism that increases the chance of the community achieving the optimal degree of diversity. In all of these models mechanisms that generate transient diversity function by preventing fully connected communities from prematurely converging on a possibly wrong theory.[1]

Since all of the above ABMs are inspired by Zollman's models,[2] which represent the situation of theoretical diversity in terms of 'bandit problems', this raises the question whether the same kind of mechanisms still played a role (in the sense of generating transient diversity) if we represented scientific inquiry in a structurally different way (as e.g. Grim et al. (2013) or Borg et al. (2018) do). Moreover, whether

---

[1] Alexander (2013) presents a slightly different scenario, where the number of rivaling theories grows over time. His results suggest that some learning strategies (namely, the combination of reinforcement and social learning via preferential attachment) can lead to the optimal level of diversity, under the condition that agents discount the knowledge of past theories.

[2] We have omitted a class of models employing epistemic landscapes (such as Weisberg and Muldoon 2009; Alexander et al. 2015; Thoma 2015; Pöyhönen 2017) since they tend to represent a different kind of diversity than the one we are focusing on in this paper: they rather examine what would better be labeled as 'cognitive diversity' (Pöyhönen and Kuorikoski 2016), which concerns different research heuristics employed by individual agents across the given community. Moreover, efficiency of inquiry in these models is usually measured in terms of success of the community in discovering certain patches of the given landscape, rather than in terms of agents converging on a single theory.

transient diversity is a robust property in the sense that a certain degree of a diversity is a 'difference-making' factor when it comes to successful inquiry, is another open question.

Addressing this issue is important not only for the examination of the robustness of previous results, but also for a more precise understanding of the phenomenon of transient diversity and its relation to the efficiency of inquiry (where efficiency is a function of the rate of successful convergence and the required time).

In this paper we will examine this question by means of an argumentation-based ABM (ArgABM) of scientific interaction, which we previously presented in Borg et al. (2017).[3] We will focus on two kinds of interrelated mechanisms:

1. On the one hand, we will examine mechanisms that represent *cautious decision-making* (previously discussed by Frey and Šešelja (2018a) with respect to a Zollman-inspired model). The first such mechanism is 'rational inertia' that an agent has towards her pursued theory, which assures she abandons the theory only after having repeatedly gathered evidence in favor of its rival for a significant period of time. The second mechanism is a relative threshold value which a rivaling theory has to surpass in order to count as superior to one's current theory.

2. On the other hand, we will examine different *evaluative procedures*, in view of which scientists decide which theory to pursue and on top of which cautious decision-making is employed. For instance, agents in the model may prefer theories that have a wider scope than their rivals, or they may avoid theories that exhibit more anomalies than their rivals. These measures may come down to different preference orders on the given theories. While ABMs of scientific interaction have usually employed a specific kind of assessment, which of these assessments is either descriptively adequate or normatively desirable has largely remained open. To this end, it is helpful to understand their impact on the efficiency of inquiry.

What makes ArgABM especially suitable for this research question is that, on the one hand, it employs both of the above mechanisms representing cautious decision-making as parameters of the model. On the other hand, the model allows for a straightforward approach to studying different assessment procedures underlying the theory choice of scientists. In addition, the model employs a specific approach to knowledge representation, which is structurally different from Zollman's or Grim & Singer's models. For instance, both defensible and anomalous parts of knowledge can be located as specific parts of the given theories. This makes the model apt for the above mentioned robustness analysis.

Our results suggest that, a certain degree of diversity can be clearly identified as correlated with efficient inquiry only when agents employ a specific theory-choice assessment—namely, when they prefer theories that are based on a comparatively larger body of solidified research, relative to their rivals. In that case cautious decision-making has a positive impact on the efficiency of fully connected communities. When

---

[3]The model is programmed in NetLogo (Wilensky 1999). The code of the model employed in this paper can be found at: https://github.com/g4v4g4i/ArgABM.

it comes to other evaluations, as well as to less connected communities, cautious decision-making either has no impact on the efficiency or it is harmful for it. Hence, this study indicates that determining factors conducive to the efficiency of inquiry is highly dependent on the specific model and its idealizations. This points to an important task for future research: specifying which types of inquiry (for example, related to specific scientific domains) are more adequately represented by some of these conditions and certain ABMs of science, rather than others.

Here is how we will proceed. In Section 2 we will present the central features of ArgABM. In Section 3 we will introduce four different types of evaluation underlying scientists' decision as for which theory to pursue. In Section 4 we will explicate how we model cautious decision-making. In Section 5 we will present our results: we will show how different social networks perform in each of the four evaluations, with and without the mechanisms of cautious decision-making. Moreover, we will analyze the impact of diversity on successful inquiry. In Section 6 we will conclude the paper suggesting some questions for future research.

2 ArgABM: an overview

In this section we introduce ArgABM, an argumentation-based ABM of scientific inquiry, which has previously been used for the examination of epistemic effects of scientific interaction under different types of social networks (Borg et al. 2017, 2018). The model is designed to measure the efficiency of groups of agents in their knowledge acquisition. Knowledge acquisition is represented in terms of agents exploring a number of rivaling scientific theories, where they have to determine which theory is the best one. Efficiency of their inquiry is represented in terms of their success in converging on the best of the available theories, and in terms of time they need to achieve this convergence.[4]

A specific feature of this model is that it aims to represent argumentative dynamics among scientists who explore rivaling theories or research programs[5] and exchange arguments *pro* and *con* these theories along the way. To this end, the model represents the argumentative context underlying theories, within which scientists gather evidence for the hypotheses constituting the given theory and against the rivaling ones. Such an argumentative context is represented in terms of an *argumentative landscape*, explored by agents.

---

[4]In Borg et al. (2017, 2018) the efficiency in terms of time is measured in a slightly different way. Moreover, in Borg et al. (2017) we present an alternative, 'pluralist' measure of success, according to which agents are successful if at the end of the run the best theory isn't less populated than any of its rivals. In the current section we will try to keep technical details at the minimum. An interested reader can take a closer look at the above mentioned publications on this model.

[5]For the sake of simplicity, we use the terms 'theory' and 'research program' interchangeably in this paper.

2.1 The argumentative landscape

As mentioned above, the model represents scientific inquiry in which scientists explore their research programs, gradually fleshing them out. They do so by exploring the argumentative landscape, which represents the argumentative context underlying the rivaling research programs. Each theory is represented as consisting of a number of arguments. These arguments are represented abstractly, as nodes in a directed graph, connected via a *discovery relation*. An argument can be understood as a hypothesis supported by evidence gained by means of a certain study (e.g. an experiment).[6] The discovery relation represents paths that agents take when moving on the landscape, from one argument to another. Its role is to track the temporal aspect of research where new research steps build on the previous ones. Moreover, arguments belonging to one research program can attack arguments of one of the rivaling programs. Such an attack represents, for instance, a discovery of a methodological problem in a certain study of the rivaling research program, or results of a novel study which provides a better explanation of a certain phenomenon than a study offered within the rivaling program.[7] The landscape then consists of different argumentative rooted trees, with nodes as arguments and edges as discovery relation, where an argument in one tree may attack an argument in another tree (see Fig. 1).[8] The extent to which each research program is attacked is a parameter of the model. We represent all theories as trees of the same size, i.e. consisting of the same number of arguments.

While at the beginning of the run, agents only see the root argument of each theory, over the course of the run they gradually discover the rest of the landscape. Each argument can be understood as a hypothesis investigated by scientists. Throughout their exploration of the landscape, our scientists will occasionally encounter

---

[6]For a concrete example of a scientific controversy—namely, the continental drift debate—represented by means of a similar framework (based on abstract argumentation) see Šešelja and Straßer (2013).

[7]While we can imagine a situation in which a single argument serves as an objection attacking the rivaling theory in whole (for example, showing the theory cannot explain a certain set of phenomena) in the current model we abstract away from such cases by employing the idealization that attacks always target a specific part of a theory (e.g. an attack on a study in a rivaling research program pointing to a methodological problem doesn't necessarily attack results of other studies within the same program—i.e. other arguments). Note that this is already a step further in the direction of representational adequacy in comparison to Zollman-inspired ABMs. It remains a task for future research to examine whether our results remain robust if we implemented a more detailed representation of argumentative attacks, e.g. by introducing an explanatory relation between arguments and a set of explananda (as it is done by Šešelja and Straßer 2013) and more refined evaluation procedures (as compared to the ones to be introduced in Section 3).

[8]The representation of our landscape is inspired by abstract argumentation frameworks (Dung 1995). Formally, the landscape is given by a triple $\langle \mathcal{A}, \rightsquigarrow, \hookrightarrow \rangle$ where $\hookrightarrow$ is the discovery relation, $\rightsquigarrow$ is the attack relation, and $\mathcal{A} = \langle \mathcal{A}_1, \ldots, \mathcal{A}_m \rangle$ is partitioned in m many theories $T_i = \langle \mathcal{A}_i, a_i, \hookrightarrow \rangle$ which are trees with $a_i \in \mathcal{A}_i$ as a root and

$$\rightsquigarrow \subseteq \bigcup_{\substack{1 \leq i, j \leq m \\ i \neq j}} (\mathcal{A}_i \times \mathcal{A}_j) \quad \text{and} \quad \hookrightarrow \subseteq \bigcup_{1 \leq i \leq m} (\mathcal{A}_i \times \mathcal{A}_i).$$

Specifying $\rightsquigarrow$ like this ensures that the theories are conflict-free, i.e. that there are no attacks between the arguments of the same theory.
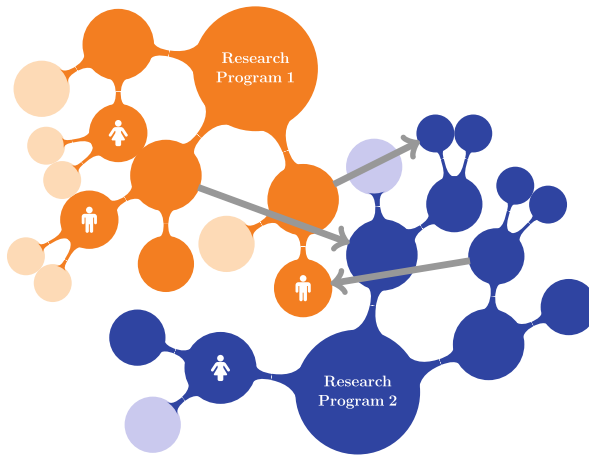
Fig. 1 An example of an argumentative landscape consisting of 2 theories (or research programs). Darker shaded nodes represent arguments that have been investigated by agents and are thus visible to them; brighter shaded nodes stand for arguments that aren't visible to agents. The biggest node in each theory is the root argument, from which agents start their exploration via the discovery relation, which connects arguments within one theory. Arrows stand for attacks from an argument in one theory to an argument in another theory

defeating evidence, represented as attacks coming from arguments in a rivaling theory. Moreover, they may encounter arguments that *defend* their attacked hypotheses, where—informally speaking—an argument a is defended in the theory if it is not attacked or if each attacker b from another theory is itself attacked by some defended argument c in the current theory.[9]

Let's look at the example illustrated in Fig. 2. In graph (a) we have argument a_1 from theory T_1, which is attacked by argument b_1 from theory T_2. In this case, a_2 defends a_1 since it attacks b_1, the attacker of a_1. If in the further course of exploration agents encounter b_2, which attacks a_2 (graph (b)), then the previous defense becomes unsuccessful and both a_1 and a_2 will now be *undefended* (for a formally precise definition of defended arguments see below Section 3.1).

The idea behind such argumentative dynamics stems from the defeasible character of scientific reasoning, where throughout inquiry scientists may encounter defeating evidence for their previously accepted hypotheses, and evidence in support of hypotheses that they have earlier rejected. This feature allows for the representation of errors that commonly appear in scientific research: false positives (accepting a

---

[9]Agents discover attacks to and from their current arguments, as well as the child arguments of their current arguments gradually, depending on the *degree of exploration* assigned to the current argument at a given time point of a run: for each agent ag and each argument $a \in \mathcal{A}$, $\mathfrak{expl}(a, \mathsf{ag}) \in \{0, \dots, 6\}$ where 0 indicates that the argument is unknown to ag and 6 indicates that the argument is fully explored and cannot be further explored. Since the model is round-based, each round may be interpreted as one research day. Each of the 6 levels of an argument takes a researcher 5 rounds/days of exploration. Thus, each argument represents a hypothesis that needs altogether 30 research days to be fully investigated.
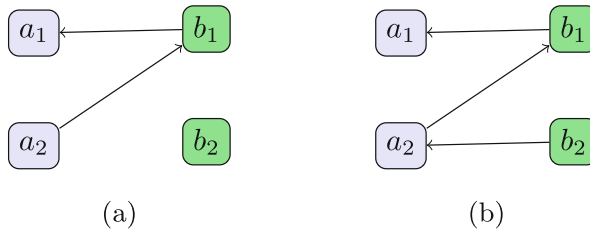
Fig. 2 Argumentation graph with arguments a_1 and a_2 belonging to theory T_1, and arguments b_1 and b_2 belonging to theory T_2. Discovery relations are omitted

hypothesis that is actually false) and false negatives (rejecting a hypothesis that is actually true). This is important in a model that aims to examine the efficiency of scientific inquiry, since these errors have a direct impact on it. Cases in which scientists accepted a false hypothesis (sometimes simulateously with rejecting a true one) are well known from the history of science.[10] This is precisely why Zollman-inspired models examine the efficiency of inquiry by focusing on the mechanisms that are conducive to minimizing the risk of false positives and false negatives.

The argumentative dynamics in our model allows for a straightforward representation of false positives and false negatives: the former are arguments that initially appear defensible, though further inquiry would reveal that they are not; the latter are arguments that are attacked and undefended, but for which a defense can eventually be found.

Now, an important feature of the model is that one of the rivaling research programs is designed as the 'best one'. In this way we can measure the efficiency of scientists by assessing their success and time needed to converge on this particular theory. The best theory is simply the one which is designed as fully defended from all the attacks, in the fully explored landscape.[11] This is, of course, an idealization, but it helps to represent the above mentioned appearance of false positives and false negatives: while at early stages of inquiry, the best theory may appear to have many anomalies (undefended arguments), if scientists keep on exploring it, they will find solutions for these anomalies (namely, defenses of the attacked arguments).

2.2 Behavior of agents

The model is round-based and each round agents perform one of the following actions:

A1   exploring a single argument, thereby gradually discovering possible attacks (on it, and from it to arguments that belong to other theories) as well as discovery relations to neighboring arguments;

---

[10]For instance, the continental drift debate or the research on peptic ulcer disease are some of the cases in point (see Šešelja and Weber 2012; Šešelja and Straßer 2014b).

[11]The other theories are modeled as having a certain percentage of their arguments attacked and undefended.

A2   moving to a neighboring argument along the discovery relation within the same theory;

A3   moving to an argument of a rivaling theory.

As mentioned above, agents start the run of the simulation at the root of a given theory and gradually discover more and more of the argumentative landscape. In this way each turn an agent operates on the basis of her own (subjective) fragment of the landscape, which consists of arguments that she has explored to a specific degree, and (attack and discovery) relations that she has found between the arguments.

To decide whether to keep on pursuing their current theory (actions A1 and A2 above), or whether to better start working on an alternative theory (A3) agents are equipped with the ability to evaluate theories.[12] Every few rounds they apply an evaluative procedure, with respect to the set of arguments and attacks they currently know (i.e. their subjective memory). We will introduce four such procedures in the next section. For now, it will suffice to say that all such evaluations are based on the question, how many defended or undefended arguments the theory has.

2.3 Social networks

Just like other models of scientific interaction, ArgABM employs social networks. In particular, agents form their subjective knowledge of the landscape not only in view of information they gather on their own, but also in view of information they receive from other agents, with whom they are linked in a network. There are two types of such networks:

1.   Collaborative groups, which consist of five individuals who start from the same theory. While each agent gathers information about the landscape on her own, every five time steps this information is shared with all other agents forming the same collaborative network.

2.   Community-networks, between collaborative groups, which are formed out of representative agents from each of the linked collaborative networks (one representative agent per collaborative network). Within community-networks agents share information (arguments and attack relations) that they have recently gathered via their exploration. This could be interpreted as having a scientist report on her recent (positive and negative) findings concerning her current theory, by writing a paper or giving a conference talk.[13] Community-networks can have one of the following structures: a *cycle*, in which each collaborative group is connected to exactly two other groups, a *wheel* which is similar to the cycle, except

---

[12]In addition, agents are equipped with a certain heuristic behavior, which allows them to search for the defense of their current argument in case it is attacked. See Borg et al. (2017), Section 2.2.

[13]In the current model we assume that agents reliably share information, i.e. that they share both positive and negative findings about their current theory. Borg et al. (2018) examine in addition deceptive information sharing, i.e. agents who share only positive findings about their theory (arguments and attacks to other theories), while withholding the information about attacks on their own theory. Whether the results presented in this paper also hold for deceptive agents remains a question for future research.
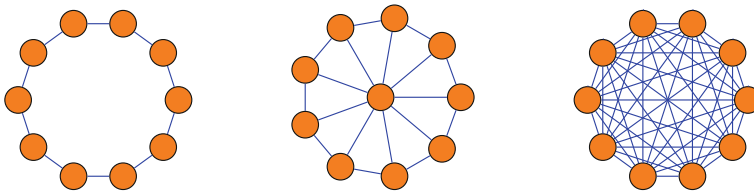
Fig. 3 A cycle, a wheel and a complete graph. Each node is a collaborative group, while the edges represent communication channels

that a unique group is connected to every other group, and a *complete graph* where each group is connected to all other groups (see Fig. 3).

3 Evaluations underlying theory-choice

As mentioned in the previous section, agents in ArgABM assess their theory in order to decide whether to stick with it, or to switch to one of the rivaling theories. In this section we will present four evaluative procedures, in view of which scientists can make such a theory-choice.

In order to explore the space of possibilities, we start with two simple measures, and then proceed by adjusting them towards two additional, more complex measures. Of course, which of these measures (or yet some other ones) is actually employed by scientists is an empirical question, which cannot be answered from a philosophical armchair.

We will motivate four suggestions for implementing such evaluative procedures in the context of ArgABM (see Section 6 for some additional proposals).

3.1 The degree of defensibility (assessment D)

Our first measure is the assessment of theories in terms of their degree of defensibility.[14] We will call it for short: *assessment D*. The degree of defensibility of a theory is the number of defended arguments in this theory. T_1 is preferred to T_2 *iff* T_1 has more defended arguments than T_2.

This strategy represents scientists who are easily impressed by the size of a theory, that is, by the size of its defensible parts.[15] In other words, they keep on pursuing their current theory unless one of the rivaling theories turns out to have more defended arguments.

---

[14]This measure is employed in previous versions of ArgABM (Borg et al. 2017, 2018).

[15]An alternative way to interpret this assessment is in terms of an explanatory scope of a theory, where we are assuming that the arguments constituting the given theory are explanatory in nature (see Šešelja and Straßer 2013). A less idealized measure of explanatory power could be implemented by introducing a set of explananda E and an explanatory relation from some of the arguments in the theory to a subset of E.
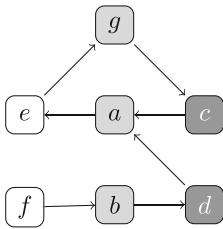
theory	defended	degree of def.
$T_1 = \{e, f\}$ (white)	$\{f\}$	1
$T_2 = \{a, b, g\}$ (gray)	$\{\}$	0
$T_3 = \{c, d\}$ (dark gray)	$\{\}$	0

Fig. 4 Argumentation Framework 1

Let's give a more precise formal definition. First, we call a subset of arguments A of a given theory T *admissible* iff for each attacker b of some a in A there is an a' in A that attacks b. Since every theory is conflict-free, it can easily be shown that for each theory T there is a unique maximally admissible subset of T (with respect to set inclusion). An argument a in T is said to be *defended in T* iff it is a member of this maximally admissible subset of T.[16] The *degree of defensibility* of T is equal to the number of defended arguments in T.

Figure 4 depicts a situation with three theories as it might occur from the perspective of a given agent: T_1 consisting of arguments e and f (white nodes), T_2 consisting of arguments a, b and g (gray nodes), and T_3 consisting of arguments c and d (dark gray nodes). The arrows represent attacks, we omit discovery relations. We are now interested in the degrees of defensibility our agent would ascribe to the given theories. The table shows which arguments are defended in each theory and their corresponding degree of defensibility. The only defended argument in this situation is f in theory T_1. Note for instance that in T_3 the argument d is not defended since no argument *in T_3* is able to defend it from the attack by b. Although the argument f in T_1 attacks b, it doesn't count as a defender of d for theory T_3 when determining the defended arguments in T_3 since in our account a theory is supposed to defend itself.

Figure 5 depicts the situation after an attack from a to f has been discovered. Consider theory T_2. In this situation a defends b from the attack by f, b defends a from the attack by d, a defends g from the attack by e and g defends a from the attack by c. Hence, all arguments are defended resulting in a degree of defensibility of 3.

3.2 The degree of anomaly (assessment A)

According to this measure, T_1 is preferred to T_2 *iff* T_2 has more undefended arguments than T_1. We call it for short: *assessment A*. If we interpret the number of undefended arguments as *the degree of anomaly* of the given research program, this strategy can be taken as representing scientists who abandon theories that become more anomalous than their rivals. This approach could be seen as corresponding to a

---

[16]Given that theories in the model are conflict-free, the notion of admissibility is here the same as the one introduced in Dung (1995). In Dung's terminology, our sets of defended arguments correspond to preferred extension (which are exactly the maximally admissible sets), except that we determine these sets relative to given theories.
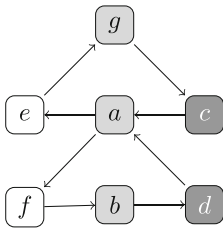
theory	defended	degree of def.
$T_1 = \{e, f\}$ (white)	$\{\}$	0
$T_2 = \{a, b, g\}$ (gray)	$\{a, b, g\}$	3
$T_3 = \{c, d\}$ (dark gray)	$\{\}$	0

Fig. 5 Argumentation Framework 2

Kuhnian scientist who resists converting to a new paradigm until her theory is clearly more anomalous than its rival (see Kuhn 1962).

Taking a look at the scenario in Fig. 4, T_1 has a degree of anomaly 1, while T_2 has a degree of anomaly 3 and T_3 has a degree of anomaly 2. Hence, agents will prefer T_1. In Fig. 5 T_1 and T_3 have a degree of anomaly 2, while T_2 has a degree of 0. Here they will thus prefer T_2.

3.3 Multiplication (assessment M)

We now turn to more sophisticated assessments. According to the measure which we call 'multiplication', T_1 is preferred to T_2 *iff* $|\text{Undef}(T_1)| \cdot |\text{Disc}(T_1)| < |\text{Undef}(T_2)| \cdot |\text{Disc}(T_2)|$, where $\text{Undef}(T_i)$ stands for undefended arguments of theory T_i, and $\text{Disc}(T_i)$ stands for all discovered arguments of T_i (i.e. arguments that belong to the knowledge base of the agent). We call this procedure for short: *assessment M*.

This strategy represents scientists who are less forgiving toward anomalies in their research program the more advanced it is (i.e. the more arguments it has). This approach could be seen as corresponding to the Lakatosian idea that in their early stages research programs are infested with anomalies, which are expected to be resolved as time passes by (see Lakatos 1978).

Taking a look at the example in Fig. 4, if we assume all the arguments in the framework are actually discovered, then T_1 has a multiplication score of $1 \times 2 = 2$, T_2 has a score of $3 \times 3 = 9$ and T_3 has $2 \times 2 = 4$. Agents will thus prefer T_1.

3.4 Normalization (assessment N)

Our final measure is labeled 'normalization' since according to it, T_1 is preferred to T_2 *iff* $|\text{Undef}(T_1)|/|\text{Disc}(T_1)| < |\text{Undef}(T_2)|/|\text{Disc}(T_2)|$, where again $\text{Undef}(T_i)$ stands for undefended arguments of theory T_i, and $\text{Disc}(T_i)$ stands for all discovered arguments of T_i.[17] We call this evaluation for short: *assessment N*.

This strategy represents scientists who evaluate the defended (or anomalous) scope of their research program relative to how advanced it is. The idea behind this assessment is similar to Bayesian updating via beta-distributions (employed by Lakatos

---

[17]It is easy to show that the following measure results in an equivalent preference order: T_1 is preferred to T_2 *iff* $|\text{Def}(T_1)|/|\text{Disc}(T_1)| > |\text{Def}(T_2)|/|\text{Disc}(T_2)|$, where $\text{Def}(T_i)$ stands for defended arguments.

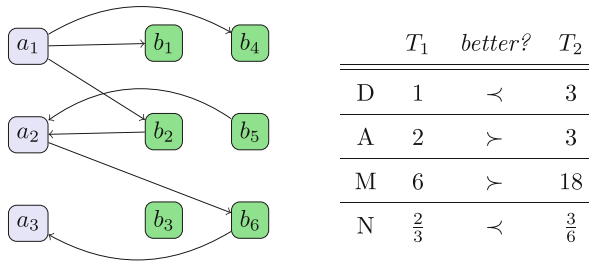|	T_1	better?	T_2
D	1	$\prec$	3
A	2	$\succ$	3
M	6	$\succ$	18
N	$\frac{2}{3}$	$\prec$	$\frac{3}{6}$

Fig. 6 Argumentation framework illustrating different evaluations underlying theory-choice. D: degree of defensibility, A: degree of anomaly, M: multiplication, N: normalization

2010), the mean of which is given by the ratio of the number of successful draws from the distribution through the number of all draws.

Considering the example in Fig. 4 and assuming all the arguments are discovered, the normalization score for T_1 is $1/2 = 0.5$, for T_2 it is $3/3 = 1$, and for T_3 it is $2/2 = 1$. Thus, agents prefer T_1.

While applying our four evaluations to the example in Fig. 4 has led to the same preference order (with T_1 being selected in each case), the following example illustrates that our four assessments may not always lead to the same theory-choice.

The example in Fig. 6 consists of two theories, a blue one, T_1, with arguments a_1-a_3, and a green one, T_2, with arguments b_1-b_6. We have that $\text{Disc}(T_1) = 3$, $\text{Def}(T_1) = 1$, $\text{Undef}(T_1) = 2$, $\text{Disc}(T_2) = 6$, $\text{Def}(T_2) = 3$ and $\text{Undef}(T_2) = 3$. Hence, T_1 has a multiplication score of 6 and a normalization score of $\frac{2}{3}$ and T_2 has a multiplication score of 18 and a normalization score of $\frac{3}{6}$. Therefore, T_1 is preferred over T_2 if theories are compared by means of assessments A or M, and T_2 is preferred over T_1 when evaluation is done by means of assessments D or N.

4 Modeling cautious reasoning

We will now explicate two types of diversity-preserving mechanisms, each of which can be understood in terms of cautious reasoning, that functions in combination with evaluations presented in the previous section.

4.1 Rational inertia: temporal threshold

The first mechanism has the aim to prevent agents from being hastily swayed by new evidence. It functions in the following way: an agent abandons her current theory and switches to a rivaling one only after she has received consistent evidence showing that the latter is better for X number of evaluations (where X is a parameter of the model). We will refer to X as *temporal threshold*. This corresponds to the situation in which scientists don't easily abandon their theory, even after discovering problems with it. Instead, they stick with it until and unless they are convinced that it can no longer be saved from the defeating evidence and that its rival is superior to it.

We call such an inertia *rational* for it wouldn't make much sense for a scientist to prematurely abandon her theory, before she is sure the current anomalies cannot be resolved and the theory improved. In this sense, it is rational for a scientist to stick to her theory for a while longer (see Kelp and Douven 2012[18]). Moreover, such an inertia is rational also in view of the fact that changing one's inquiry usually comes with various costs (such as acquiring additional knowledge, new equipment, etc.).

4.2 Similarly successful theories count as equally good: epistemic threshold

While a rational inertia keeps agents 'sticky' on their theories for a certain period of time, our second mechanism keeps them 'sticky' for as long as the rivaling theory isn't significantly better than their current one. To this end, agents stay on their current theory unless it has been surpassed by a rival beyond a given threshold value, relative to the employed evaluation procedure. We call such a threshold – *epistemic threshold*.

More precisely, an agent abandons her current theory only if it fails to be one of the best theories, where the set of 'best theories' is calculated by means of the four evaluative procedures together with the epistemic threshold in the following way:

- For the evaluation in terms of assessment D: if T_i stands for a theory that has the highest degree of defensibility, then the set of best theories consists of those theories that have at least the following assessment D:

$$|\text{Def}(T_i)| \cdot [\text{epistemic threshold}]$$

  where epistemic threshold is a value from the interval $(0, 1]$.
- For the evaluation in terms of assessments A, M and N: if T_i stands for a theory that has the lowest Evaluation Score(T_i) for each of the three measures, then the set of best theories consists of those theories that have maximally the following score:

$$\frac{\text{Evaluation Score}(T_i)}{[\text{epistemic threshold}]}$$

  where epistemic threshold is a value from the interval $(0,1]$.

For instance, let T_i be a theory with Disc$(T_i) = 20$, Undef$(T_i) = 10$ and Def$(T_i) = 10$ and assume T_i is the theory with the most defended arguments and the lowest evaluative score according to the A, M and N procedures. We choose the epistemic threshold of 0.9. For each of the evaluation procedures we get the following scores:

- D: all theories that have at least $10 \cdot 0.9 = 9$ defended arguments will fall among the set of best theories,

---

[18]Rational inertia shouldn't be confused though with the 'Steadfast Norm' discussed by Kelp and Douven in the same paper, and well-known in the literature on peer disagreement. Unlike in their account, in our model we may interpret a scientist as having a rational inertia towards her theory, while having lowered her confidence that the theory is actually true.

Table 1 Set of best theories for the example in Fig. 6, for different epistemic threshold values

Epistemic threshold	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$
D	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_2\}$	$\{T_2\}$	$\{T_2\}$
A	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1\}$
M	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1\}$	$\{T_1\}$	$\{T_1\}$
N	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1, T_2\}$	$\{T_1, T_2\}$

- A: all theories whose degree of anomaly is smaller than $10/0.9 = 11.11$ count among the best ones,
- M: all theories whose multiplication score is less that $(10 \cdot 20)/0.9 = 222.22$ count among the best ones,
- N: all theories whose normalization score is less than $(10/20)/0.9 = 0.55$ count among the best ones.

The primary idea behind this mechanism is that a rivaling theory has to pass a sufficiently wide margin to be considered superior to one's current theory. This corresponds to the reasoning of a scientist who uses a dose of caution in such evaluations, knowing that future inquiry might reveal new evidence. As a result, she will abandon her current theory not merely after she has seen it perform worse a multiple number of times (as in the case of rational inertia), but only after its rival has become sufficiently superior to it.

In Table 1, we show the sets of best theories for the example in Fig. 6, for different values of epistemic thresholds.

5 Our findings

In this section we present the results of our simulations, focusing on two measures: how successful agents are in converging on the best theory, and how much time they need to converge on it.[19] Each of the plots shows a mean of 10,000 simulations for each data point (unless otherwise indicated). All the simulations were run with a landscape consisting of 3 theories, each having 85 arguments. While the best theory is fully defended, the other two theories have a certain portion of undefended arguments.

Concerning the last point, we employ two types of landscapes:

1. an 'easy' landscape, in which the two suboptimal theories have around 35% of undefended arguments,[20]

---

[19] Due to space restructions, many of the plots are omitted from the paper and can be found in the Online Appendix of this paper.

[20] For the exact procedure of how the attacks are generated, and the degree of defensibility of the two suboptimal theories such a procedure results in, see Borg et al. (2018).

2. a 'difficult' landscape, in which the two suboptimal theories have around 85% of undefended arguments.

That a landscape is easy/difficult means that theories are more or less similar in terms of their degree of defensibility, which makes it easier or harder to determine which one is the best.

A simulation stops when one of the theories is fully explored. At this point we examine whether the agents have converged on the best theory, and if so, at which step of the simulation they have done so.[21]

As for our two mechanisms explicated in the previous section—which we call for short 'threshold mechanisms' or 'thresholds'—we have employed the temporal threshold of 10. This means that in order for an agent to switch to a rivaling theory, she has to consistently evaluate that theory as one of the best ones (and better than her current theory) for 10 (not necessarily consecutive) rounds.[22] For the epistemic threshold, we have opted for a relatively small value of 0.9. We have tested our results with higher thresholds (e.g. temporal threshold of 50, and the epistemic threshold of 0.7) and they have remained robust under these changes, except for the time agents need to achieve convergence, which, as expected, increases with higher thresholds.

5.1 Results

We will now focus on four interesting points revealed by the simulations. In the next subsection we will discuss these findings.

Impact of threshold-mechanisms First, the impact of the threshold mechanisms varies across different evaluative procedures. The only case where we observe a positive effect of thresholds on the success of agents is the complete graph employing procedure D. The impact of thresholds on different networks employing assessment D can nicely be observed in case of a larger population (of 70 agents), represented in Fig. 7. In case of all other evaluations and network structures, thresholds either have no effect or they have a negative effect, across both easy and difficult landscapes (see Table 2).

Efficiency of different evaluative procedures Second, different evaluative procedures result in drastically different degrees of efficiency, across all three networks. While D assessment results in the worst performance for all three networks in case of both types of landscapes, N procedure makes all three networks very efficient on the easy

---

[21]The reason why we stop the simulation at this point is that otherwise some agents would become 'idle': since they have explored their preferred theory fully, the only way they would change their preference is by waiting for other agents to send them new information. Borg et al. (2019) propose an alternative model in which the simulation continues after this point, eventually bringing all agents on the best theory, so that the efficiency is measured in terms of time only (similarly to the ABM proposed by Frey and Šešelja 2018a).

[22]In view of an interpretation suggested by Borg et al. (2017), according to which a round stands for a working day, this threshold means that scientists have to wait 10 weeks before being able to change their theory. Of course, different interpretations of the time in the model are possible.
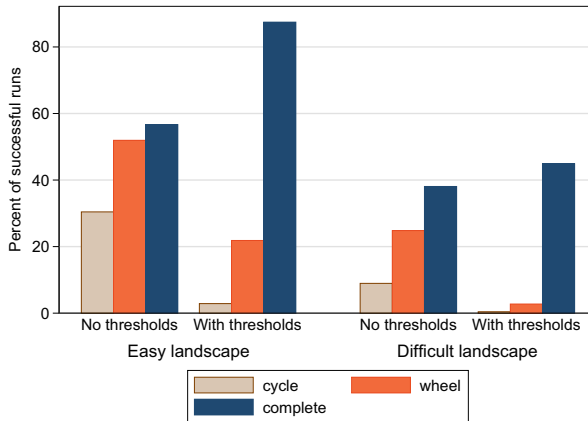
Fig. 7 Success of agents employing procedure D (70 agents)

landscape. Nevertheless, a complete graph employing the M procedure overtakes the N one on the difficult landscape (see Figs. 8 and 9).

Efficiency of different social networks Third, the relative efficiency of different social networks remains pretty robust across all explored scenarios, with the complete graph outperforming less connected networks in terms of both – the success of agents in converging on the best theory, and the amount of time they need to achieve such convergence. In the case of A, M and N evaluations the complete graph is extremely successful on the easy landscape, while being a bit less successful on the difficult one.

Transient diversity As mentioned in Section 1, the literature on ABMs of science has advanced the idea that in order to optimize efficiency of scientific inquiry we seek a diversifying mechanism that creates a tension among agents such that it is (a) strong enough to prevent agents from an early convergence on the wrong theory and (b) sufficiently soft to enable them to eventually converge on the right theory. The wanted type of diversity has been labeled *transient*. One ingredient of such diversity was

Table 2 The impact of threshold-mechanisms on different social networks with respect to the four evaluative procedures on the easy landscape (on the left) and on the difficult landscape (on the right with shaded background). Complete: complete graph; Sub-Complete: cycle and wheel networks; +: positive impact; −: negative impact; ±: neither positive nor negative impact. Note that the effect is the strongest for larger populations

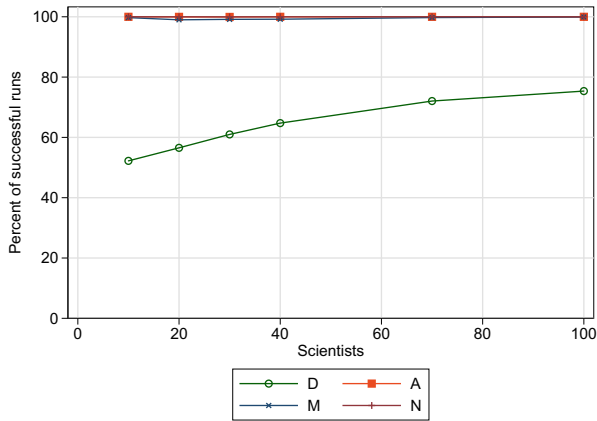|	Success				Time			
Evaluation	Complete		Sub-Complete		Complete		Sub-Complete	
D	+	+	−	−	−	−	−	−
A	±	±	±	±	±	±	±	±
M	±	±	−	−	±	±	±	±
N	±	±	±	±	−	±	−	±

Fig. 8 Easy landscape: success of agents connected in the complete graph for different evaluation procedures (aggregated over both runs with thresholds and without thresholds)

identified in the social network structure, another one in epistemic biases (Zollman 2010). In this paper we have studied other parameters, such as evaluative standards of agents and (temporal and epistemic) thresholds used by agents when deciding when to choose another theory.

Our first expectation is that higher thresholds have a diversifying effect similar to loser network structures. And indeed this is what we see for instance in Fig. 10 for the D and N procedure. We measure diversity of a run in terms of the number of rounds in which agents have no consensus on any theory divided by the number of rounds it took to terminate the run. We can see that the center of mass is moved to the right (more diversity) when introducing thresholds.
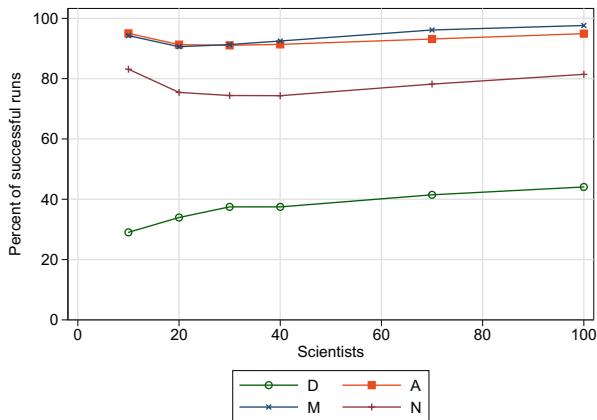

Fig. 9 Difficult landscape: success of agents connected in the complete graph for different evaluation procedures (aggregated over both runs with thresholds and without thresholds)
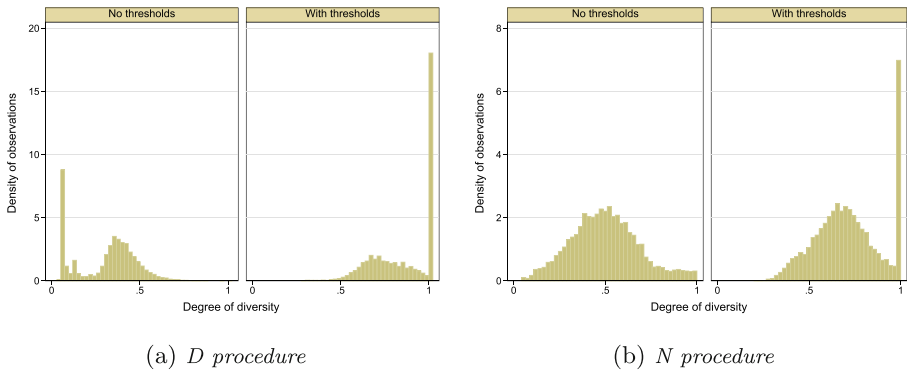
(a) *D procedure*   (b) *N procedure*

Fig. 10 The effect of thresholds on the degree of diversity (70 agents, difficult landscape, complete graph)

When considering the relation between the degree of diversity and efficiency we may naively expect a bell-shaped curve at whose peak we find runs with most efficiency while moving to more or less diversity the situation worsens. Things are more complicated, though. We find, for instance, a camel-like curve for the D-procedure and difficult landscapes (see Fig. 11) with one peak for runs with diversity degrees between 0 and 0.1, and another peak for runs with diversity degrees between 0.7 and 0.8. Furthermore, the difficulty of the landscape influences the shape of the curve: for easy landscapes more diversity is highly beneficial as we can see for the interval from 0.5 to 0.8, but less so for low diversity degrees (unlike in the difficult landscape). Also the evaluation criterion matters, as we can observe when considering the N procedure where we see a continuous (for a long time slow) decline of efficiency with higher degrees of diversity.

In sum, the efficiency-diversity relation does not in general exhibit a simple bell-like curve. Moreover, the shape of the curve is highly dependent on factors such as the underlying evaluative procedure and the difficulty of the problem. Furthermore, in some cases (like the N procedure) diversity has not much of an influence on efficiency (except for extreme degrees). This also highlights the importance of studying
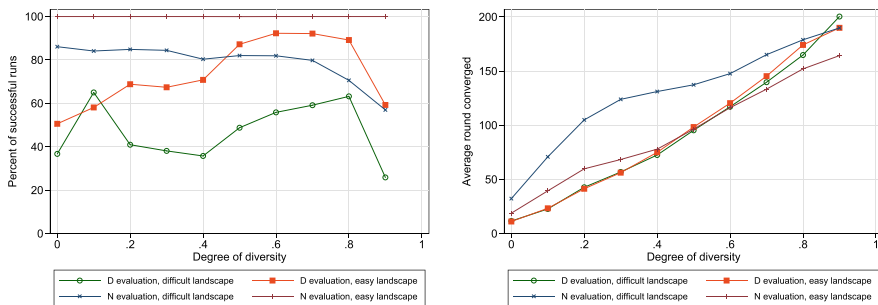
(a) *Percentage of successful runs relative to degree of diversity*   (b) *Round of convergence relative to degree of diversity*

Fig. 11 The effect of diversity on efficiency (70 agents, complete graph)

other factors which influence the efficiency of scientific inquiry, such as evaluation procedures, as done in this paper.

5.2 Discussion

We will now comment on a few most important aspects of our findings.

Highly successful communities The first striking point that deems an explanation is the extremely high success rate of fully connected communities in case of A, M and N evaluations. Why do these populations perform so well?

To answer this question, we will first explain (i) why fully connected networks tend to be at least as successful as the less connected ones, and in most cases much more successful, and then turn to (ii) the success of A, M and N evaluations in particular.

As for (i), the reason for their success lies in the way information is represented in our model. How accurate one's assessment about the given theories is, directly depends on how much knowledge of the landscape the agent has. Larger gaps in such knowledge can easily lead to errors in theory assessments. Now, since our agents share only recently acquired information (rather than their entire knowledge of the landscape), in less connected communities some of this information may easily be missed, and hence their knowledge of the landscape will be 'patchier'. As a result, they may fail to accurately determine the best theory.[23] Note that this is also why larger communities linked in sub-complete graphs have a low success rate: since in our community-networks not every agent communicates with every other agent (instead collaborative groups appoint representative agents who then share information in community-networks), the degree of connectedness gets smaller the larger the overall population is, and as a result, subjective knowledge can in larger populations be rather different across different collaborative groups. Moreover, since agents share recently gathered information, there may be a permanent information loss in such groups. This is in contrast to, e.g., Zollman's model, where any shared information is representative of the entire theory, which makes information losses much less harmful. We take ArgABM, however, to be representative of a situation in which scientists who don't share all their results may fail to have an encompassing understanding of each of the rivaling theories (e.g. they might lack an insight into an important study in one of the theories). This means that larger populations of scientists will have a harder time converging on the same theory due to the fact that they assess theories in view of different evidence. This is, however, not unrealistic: larger scientific communities that are not tightly connected indeed tend to have a harder time achieving consensus on one theory.

As for (ii), the reason why A, N and M evaluations perform better than D becomes clear when we observe that agents in the case of the former assessments tend to

---

[23]Though we haven't examined the situation in which agents share a random subset of their knowledge of the landscape (rather than only recently acquired information), the fully connected community would most likely still outperform the less connected networks since, on the one hand, it would still have a less patchier knowledge of the landscape than the other two networks, while on the other hand, such a change is not likely to increase the chance that the community prematurely abandons the best theory.
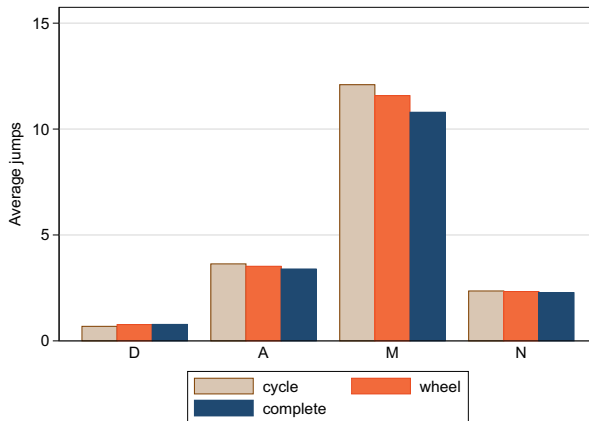
Fig. 12 The number of times agents switch from one theory to another with no threshold-mechanisms, averaged over all population sizes for the easy landscape

switch more often from one theory to another (see Figs. 12 and 13). In other words, these assessments generate diversity by allowing agents to change their theories and gain enough information about them to accurately decide which one is the best.

Cautious decision-making What do our results tell us about cautious decision-making and its conduciveness to efficient inquiry? The impact of our threshold mechanisms seems to be highly dependent on (i) the degree of connectedness of the given community, and (ii) the evaluation underlying theory choice employed by agents (as visible from Table 2). Altogether, the thresholds increase the efficiency only of fully connected communities that employ D assessment, while sometimes having the opposite effect on the less connected ones. Moreover, for A, N and M assessments the addition of thresholds just slows them down.

In view of these considerations it might seem like our mechanisms of cautious decision-making play no beneficial epistemic role at all unless scientists apply the assessment in terms of D procedure. Nevertheless, a closer look at the simulations reveals that thresholds do play an important role, which is not immediately clear when analyzing the results for success and time. Looking at the exploratory behavior of agents—how many times they switch from one theory to another—we observe that without the presence of thresholds, agents frequently switch between theories (see Figs. 12 and 13). While our model doesn't take into account that changing theories can be costly (in terms of time one needs to learn the necessary background knowledge or in terms of costs of acquiring the right equipment), in many domains this can be an important issue.[24]

---

[24]For example, different hypothesis in medicine concerning the main causes behind a given disease may require knowledge in different medical disciplines. For a further discussion on the importance of including costs of this kind into ABMs of science see Muldoon (2017).
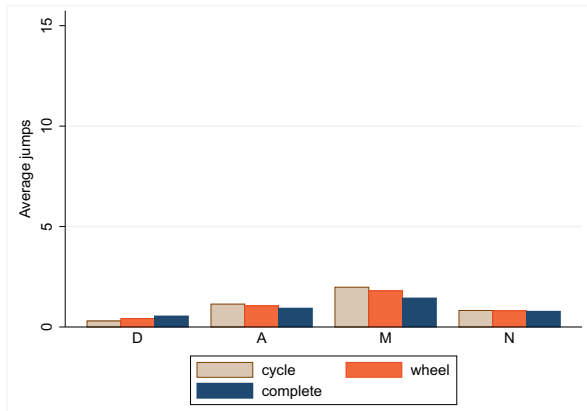
Fig. 13 The number of times agents switch from one theory to another with temporal threshold of 10 and epistemic threshold of 0.9, averaged over all population sizes for the easy landscape

This brings us to the following conclusion: while in view of previous ABMs (such as Frey and Šešelja 2018a), it seemed that threshold mechanisms played an important role in generating transient diversity in fully connected communities, our results indicate that this is the case only under certain conditions. More precisely, threshold mechanisms will have a beneficial impact only if the costs of changing theories, occurring in the absence of cautious decision-making, are high enough to make incautious communities slower than the cautious ones. This points to the importance of including this factor in ABMs of scientific inquiry. Note, however, that a proper study of such costs would require empirical calibration of the given model. First, the time in the model would have to be mapped to the real time of inquiry, and second, the costs associated with changing one's theory would have to be based on empirical data concerning the given domain of science.

The role of diversity Let's take now a closer look at the D procedure to get a better understanding of the role diversity plays in our simulations. As we can observe in Fig. 10a, without thresholds the majority of the runs is roughly located between diversity degrees 0 and 0.5 while with thresholds it is roughly between 0.5 and 1. When introducing thresholds we only get a slight increase in successful runs for the difficult landscape despite the vast difference in diversity (see Fig. 7). How to explain this? The answer is given in Fig. 11a. Given the information from Fig. 10a, we notice that without thresholds many successful runs will be located around the steep peak at 0.1 and not many around the peak at 0.7 to 0.8. When introducing thresholds the situation is exactly vice versa. Since overall the area between 0.5 and 0.8 is more elevated as compared to the area from 0 to 0.5 we get a slight boost in efficiency, but not too much.

This analysis demonstrates that when analyzing the given dynamics in our runs, diversity has explanatory value: only by combining the data given in Figs. 10a and 11a we were able to explain the only slight performance boost in Fig. 7. Nevertheless, we consider the investigations into diversity in this section preliminary for

several reasons. For instance, our way of measuring diversity is still very rough. A more refined approach may provide measures that distinguish between synchronic and diachronic diversity: the former concerns the distribution of agents among different theories at a given time point, the latter concerns the number of times agents change theories over the course of a run. Our current measure can be considered as a rough way of measuring the former. We postpone a more in-depth analysis for future work.

More general take-home message More generally, our results show that determining the impact of a specific factor on the efficiency of scientific inquiry is highly dependent on the specific model and its idealizations. While in Zollman-inspired ABMs threshold mechanisms had a big impact, in ArgABM they do so only under very specific conditions. In the former, their main role is in preventing the community from prematurely converging on the wrong hypothesis by allowing for more data to be gathered before the decision is made. This is also the case in ArgABM when agents employ D assessment on the easy landscape.

However, a much more efficient approach to increasing the efficiency seems to lie in the type of assessment underlying scientists' decisions as for which theory to pursue. Altogether, our analysis provides further support to the argument that ABMs of science are in need of detailed robustness analysis before we can draw from them any conclusions about actual scientific practice.[25]

It is also worth noticing that differences between our procedures for theory-choice could be understood as representing specific epistemic and methodological values preferred by scientists. While such preferences are still highly idealized across ABMs of science, our results suggest that methodological values may play an important role in the efficiency of inquiry and that they deserve further attention. Beside Weisberg and Muldoon's (2009) 'mavericks' and 'followers', or Currie and Avin's (2018) 'obligates' and 'omnivores'[26], other types of methodological preferences could be considered: for instance, a method based on the search for defeaters vs. a method that prioritizes corroborating evidence for one's current hypothesis, etc.

Another important take-home message is that some relevant factors may very well remain hidden unless we take an in-depth analysis of the given simulations. For instance, while the impact of the threshold-mechanisms seemed rather neutral or even harmful for three of our evaluations, only once we have examined how often scientists change theories, it has become obvious that they did play an important role—by reducing possibly high costs that may be involved in a scientist's frequent change of a pursued theory.

---

[25]For the importance of robustness analysis for models in general see e.g. Weisberg (2006) and for ABMs of science in particular see Frey and Šešelja (2018a), Frey and Šešelja (2018b), and Šešelja (2019).

[26]While maverics and followers stand for more or less epistemically risk-averse agents, omnivorse are agents that prioritize *independent* evidence for their hypotheses, i.e. evidence that is supported by background theories that overlap as little as possible. Obligates, on the other hand, seek *sharp* evidence that "speaks clearly and firmly" (p.5): the sharper the evidence the more it allows us to increase our credence in a given hypothesis.

6 Outlook and conclusion

In this paper we have investigated the impact of different factors on the efficiency of scientific inquiry by means of ArgABM. To this end, we have examined the impact of cautious decision-making, different assessments underlying theory-choice, and different network structures on the efficiency of inquiry. In addition, we have examined the phenomenon of transient diversity by studying the relationship between a diverse, non-consensual spread of scientists across different theories and their performance under varying conditions. Our results suggest that, on the one hand, cautious decision-making has a significant impact on the efficiency of inquiry only under specific conditions. On the other hand, different assessments underlying theory-choice and different network structures result in varying degrees of efficiency. Moreover, diversity is not always correlated with a successful performance of scientists, but only under some conditions. Such a correlation occurs when scientists prefer theories that have a relatively larger scope of solidified results, in comparison to their rivals.

It is important to add though that the nature of this model and our results are primarily exploratory (rather than having normative consequences for actual scientific inquiry). The next step in this investigation includes, for instance, examining the performance of other evaluation procedures, which include the measure of the growth of the given research program.[27] Next, it would be valuable to relate these evaluations with philosophical and historical accounts of decision-making in the context of pursuit (such as Whitt 1992; Nickles 2006; Šešelja and Straßer 2014a), as well as to empirically calibrate different aspects of the model (such as the time of inquiry, the degree of anomaly of given theories, etc.). Furthermore, it remains a task for future research to determine which types of inquiry (e.g. more related to some scientific domains rather than others) are more adequately captured by Zollman-inspired models, which by Grim & Singer's ones, and which by ArgABM. Finally, our results point to the importance of further studies of the phenomenon of transient diversity and its relation to efficient inquiry.

References

Alexander, J.M. (2013). Preferential attachment and the search for successful theories. *Philosophy of Science*, *80*(5), 769–782.

Alexander, J.M., Himmelreich, J., Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, *82*(3), 424–453.

---

[27] Such an approach could capture, for instance, Laudan's (1977) suggestion that "*it is always rational to pursue any research tradition which has a higher rate of progress than its rivals* (even if the former has a lower problem-solving effectiveness)" (p. 111, italics in original).

Borg, A.M., Frey, D., Šešelja, D., Straßer, C. (2017). Examining network effects in an argumentative agent-based model of scientific inquiry. In Baltag, A., Seligman, J., Yamada, T. (Eds.) *Proceedings Logic, rationality, and interaction: 6th international workshop, LORI 2017, Sapporo, Japan, September 11-14, 2017* (pp. 391–406). Berlin: Springer Berlin Heidelberg.

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.

Borg, A.M., Frey, D., Šešelja, D., Straßer, C. (2019). Using agent-based models to explain past scientific episodes: towards robust fndings. *Forthcoming*.

Currie, A., & Avin, S. (2018). Method pluralism, method mismatch and method bias. *Philosopher's Imprint*.

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*, 321–358.

Frey, D., & Šešelja, D. (2018a). Robustness and idealization in agent-based models of scientific interaction. *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy039.

Frey, D., & Šešelja, D. (2018b). What is the epistemic function of highly idealized agent-based models of scientific inquiry? *Philosophy of the Social Sciences*. https://doi.org/10.1177/0048393118767085.

Grim, P. (2009). Threshold phenomena in epistemic networks. In *AAAI fall symposium: complex adaptive systems and the threshold effect* (pp. 53–60).

Grim, P., Singer, D.J., Fisher, S., Bramson, A., Berger, W.J., Reade, C., Flocken, C., Sales, A. (2013). Scientific networks on data landscapes: question difficulty, epistemic success, and convergence. *Episteme*, *10*(4), 441–464.

Kelp, C., & Douven, I. (2012). Sustaining a rational disagreement. *EPSA philosophy of science: Amsterdam 2009* 101–110.

Kuhn, T. (1962). *Structure of scientific revolutions*, 3rd edition. Chicago: The University of Chicago Press.

Kummerfeld, E., & Zollman, K.J.S. (2016). Conservatism and the scientific state of nature. *The British Journal for the Philosophy of Science*, *67*(4), 1057–1076.

Lakatos, I. (1978). *The methodology of scientific research programmes. Philosophical papers. Volume I, Editors: John Worrall and Gregory Currie*. Cambridge: Cambridge University Press.

Laudan, L. (1977). *Progress and its problems: towards a theory of scientific growth*. London: Routledge and Kegan Paul Ltd.

Muldoon, R. (2017). Diversity, rationality and the division of cognitive labor. In *Scientific collaboration and collective knowledge: New Essays. Oxford University Press*.

Nickles, T. (2006). Heuristic appraisal: context of discovery or justification? In *Revisiting discovery and justification: Historical and philosophical perspectives on the context distinction* (pp. 159–182).

Pöyhönen, S. (2017). Value of cognitive diversity in science. *Synthese*, *194*(11), 4519–4540.

Pöyhönen, S., & Kuorikoski, J. (2016). Modeling epistemic communities. In Fricker, M., Graham, P.J., Henderson, D., Pedersen, N., Wyatt, J. (Eds.) *The routledge handbook of social epistemology (forthcoming)*. Routledge.

Šešelja, D. (2019). Some lessons from simulations of scientific disagreements, synthese (accepted for publication).

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

Šešelja, D., & Straßer, C. (2014a). Epistemic justification in the context of pursuit: a coherentist approach. *Synthese*, *191*(13), 3111–3141.

Šešelja, D., & Straßer, C. (2014b). Heuristic reevaluation of the bacterial hypothesis of peptic ulcer disease in the 1950s. *Acta Biotheoretica*, *62*, 429–454.

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.

Thoma, J. (2015). The epistemic division of labor revisited. *Philosophy of Science*, *82*(3), 454–472.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730–742.

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, *76*(2), 225–252.

Whitt, L.A. (1992). Indices of theory promise. *Philosophy of Science*, *59*, 612–634.

Wilensky, U. (1999). Netlogo. (http://ccl.northwestern.edu/netlogo/). In *Center for connected learning and computer based modeling.* Northwestern University.

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

Zollman, K.J.S. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, *72*(1), 17–35.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

AnneMarie Borg[1] · Daniel Frey[2] · Dunja Šešelja[3] · Christian Straßer[1]

[1] Institute for Philosophy II, Ruhr-University Bochum, Bochum, Germany

[2] Faculty of Economics and Social Sciences, Heidelberg University, Heidelberg, Germany

[3] Munich Center for Mathematical Philosophy, LMU Munich, Munich, Germany