D6.2 Scenario modelling of policy interventions regarding data sharing for RRI and Open Science transition

on merrit

Observing and Negating Matthew Effects in Responsible Research and Innovation Transition

D6.2 Scenario modelling of policy interventions regarding data sharing for RRI and Open Science transition
Version 1.0
STATUS Public

This deliverable presents results from building and analysing the DASH model, an agent-based model that investigates how interventions at the level of research funders might lead to increased data sharing among academic research groups.

ON-MERRIT - Grant Agreement 824612

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 824612.

Document Description

D6.2 Scenario modelling of policy interventions regarding data sharing for RRI and Open Science transition

D6.2 Scenario modelling of policy interventions regarding data sharing for RRI and Open Science transition			
WP6 - Synthesis, validation and policy recommendations			
Due date	31.01.2022	Actual delivery date:	31.01.2022
Nature of document	[Report]	Version	1.0
Dissemination level	Public		
Lead Partner for deliverable	KNOW		
Authors	Thomas Klebel, Tony Ross-Hellauer		
Reviewers	Birgit Schmidt, Petr Knoth		

Revision History

Issue	Item	Comments	Author/Reviewer
V 0.1	Draft version	Initial draft of literature, results	Thomas Klebel, Tony Ross-Hellauer
V 0.2	Review	Review of Intro, Lit review, Methods, Results	Birgit Schmidt
V 0.3	Updated draft	Update based on review, add Discussion	Thomas Klebel
V 0.4	Review	Review of whole document	Petr Knoth
V 0.5	Revised draft	Incorporate feedback suggested revisions	Thomas Klebel
V 0.6	Revised draft	Final review and revisions	Tony Ross-Hellauer
V 1.0	Final version	Formatting and final proofing	Thomas Klebel

Table of Contents

Tables

Figures

Abbreviations

ABM	Agent Based Modelling
DASH model	Data Sharing Model
EC	European Commission
FAIR	Findable, Accessible, Interoperable, Reusable data
ON-MERRIT	Observing and Negating Matthew Effects in Responsible Research & Innovation Transition
OS	Open Science
PI	Principal Investigator
RDM	Research Data Management
RRI	Responsible Research & Innovation
WP	Work Package

Executive summary

Over the past 28 months, ON-MERRIT has collected evidence on how Open Science reshapes scientific endeavours and their interaction with societies and industries, conducting interviews, surveys, group discussions, document analysis and the analysis of large bibliographic databases. While a host of empirical results are now available via several comprehensive deliverables[1], the ON-MERRIT consortium is conducting a final and key step: synthesizing the knowledge gained and distilling it into actionable recommendations. The final recommendations will be based on the evidence collected so far, and refined in light of highly productive discussions with domain experts. Still, assessing the recommendation's efficacy and any unintended consequences based on the relationship "*evidence about system A"* → *policy* is hard. To close the circle by investigating the relationship *policy* → "*change in system A"*, this deliverable leverages the method of Agent-Based Modelling (ABM) to assess outcomes of potential policies by modelling patterns of the uptake of Open Science (OS) under policies proposed in the research literature.

Considering a key area of OS, the sharing of research data, we develop and discuss the DASH model, which takes a first step in improving our understanding of how interventions at the funder level might succeed to increase data sharing among academic research groups. We analyse the uptake of data sharing assuming rational agents who compare their current with their past success. Our analysis suggests that reasonable incentives for data sharing, such as considering an applications history data sharing to an equal amount in funding decisions as their publication history, are not sufficient to incur wide-spread adoption of data sharing. The contingencies inherent to the (stylised) process of research (funding → research → publications → funding) prohibit agents from adjusting their behaviour.

Considering agents that learn by imitating their peers, our investigations provide two key findings: first, assuming rational agents and moderate costs of Research Data Management (RDM) (5%), substantial incentives are necessary to bring about data sharing among research groups. The magnitude of incentives has a clear effect on the speed of the uptake of data sharing. Second, reducing costs for data sharing increases the rate of uptake particularly for fields where agents exhibit heterogeneous time-orientations (some of them revising their decisions very frequently, others revising less frequently).

Our model and analysis provide a first step in analysing proposed paths of action to encourage the sharing of research data. There are manifold options for extending the model, improving its accuracy, expanding its scope, and widening the types of policies considered. Nevertheless, the DASH model provides a solid illustration of how ABM can be leveraged to inform policy-making processes and strengthens the evidence upon which the upcoming ON-MERRIT policy recommendations (Deliverable 6.4) will be based.

Acknowledgements

We greatly thank Flaminio Squazzoni and Federico Bianchi for fruitful discussions about the model.

[1] https://on-merrit.eu/results/

1. Introduction

Success in research and innovation should primarily build and depend *on merit,* on clarity of thought, innovation of ideas, and integrity of processes, rather than on external factors like personal characteristics, prior reputation or levels of resources. Responsible Research and Innovation (RRI), and especially Open Science (including Open Access to publications and research data), public participation, and gender equality, hold the promise to make scientific endeavours more inclusive, participatory, understandable, accessible and re-usable for large audiences, especially beyond the ivory towers of universities and research institutions. However, the potential for the RRI agenda to realise these promises of "inclusive and sustainable research and innovation" depends heavily on the drivers and barriers to implementation imposed by a diverse range of institutions and individuals. Making processes open will not *per se* drive wide re-use or participation unless also accompanied by the capacity (in terms of knowledge, skills, technological readiness and motivation) to do so. Such differences are further intensified by other factors like geographic location, language abilities, technological skills, educational levels and access to basic equipment (e.g., Internet access). Those in possession of such capacities benefit from an advantage, with the effect that RRI's agenda of inclusivity is put at risk by conditions of "cumulative advantage".

Over the past 28 months, ON-MERRIT has collected evidence on how Open Science reshapes scientific endeavours and their interaction with societies and industries, conducting interviews, surveys, group discussions, document analysis and the analysis of large bibliographic databases. While a host of empirical results are now available via several comprehensive deliverables[2], the ON-MERRIT consortium is conducting a final and key step: synthesizing the knowledge gained and distilling it into actionable recommendations. The final recommendations will be based on the evidence collected so far, and refined in light of highly productive discussions with domain experts. Still, assessing the recommendation's efficacy and any unintended consequences based on the relationship *"evidence about system A"* → *policy* is hard. To close the circle by investigating the relationship *policy* → *"change in system A"*, this deliverable leverages the method of Agent-Based Modelling (ABM) to assess outcomes of potential policies by modelling patterns of the uptake of Open Science (OS) under policies proposed in the research literature.

Data sharing has become an integral part of the OS agenda, and is central to an increasing number of policies. The sharing and subsequent re-use of research data is expected to lead to substantial efficiency gains across science and industry (Directorate-General for Research and Innovation, 2016). There is substantial evidence that sharing of research data leads to a citation advantage (Colavizza et al., 2020; McKiernan et al., 2016)[3]. Game-theoretical approaches (Pronk et al., 2015) have highlighted that, as with many other aspects of OS, the overall adoption and uptake of data sharing can be understood as a problem of collective action (see also Scheliga & Friesike, 2014). That is, system-level benefits can be expected when sharing is the norm, while individuals may benefit from not sharing their own data while re-using data that others have shared (i.e., free-riding).

In this deliverable, we develop and discuss the DASH (**DA**ta **Sh**aring) model, which leverages the potential of ABM to investigate how plausible scenarios of incentivising data sharing could encourage its uptake, given

[2] https://on-merrit.eu/results/

[3] But c.f. Kwon & Motohashi (2021) on how this advantage might be offset by a competition effect, where any credit associated with the shared data quickly dissipates by others building on, and therefore superseding the initially shared data.

varying scenarios of costs associated with the sharing of data, as well as different strategies employed by individual research groups to decide on whether or not to share data. The model serves as a starting ground, enabling further evaluation of policy interventions for their efficacy, and potential repercussions for equity in science.

The document is organised as follows: We first conduct an extensive review of the literature, covering challenges inherent to policy-making, the ABM framework, including how ABM can inform policy-making processes, as well as discussing recent literature on the current state of play regarding the sharing of research data. Next, we detail the model's parameters and assumption, as well as explaining the interventions we investigate. We present the results in three sections, first covering how the model behaves without data sharing (baseline scenario), and then investigating how agents react to costs and incentives for data sharing when updating their sharing decisions based on (a) their individual performance, and (b) their peer's performance. Finally, we discuss our findings with respect to the current literature, detail the model's limitation, and provide an outlook how the model could fruitfully be extended in the future.

2. Literature review

The following review of the literature serves four main aims: First, we discuss common challenges of policy-making and review studies proposing Agent-Based Modelling (ABM) as a valuable tool in this process. Second, we introduce the reader to the method of ABM, including issues of calibration and validation. Third, we review existing studies that have leveraged ABM to inform policy-making. Finally, we discuss studies investigating the extent of data sharing across disciplines, including recently suggested interventions to encourage wider adoption.

2.1. Challenges of Policy-Making

"Predictions are hazardous, especially about the future." This common proverb succinctly summarises one key challenge faced by policy-makers: today's decisions target the future, and potentially under considerable levels of uncertainty. It is for this reason that policy-making commonly seeks to reach decisions which are based on current knowledge by relying on the "state-of-the-art science and technology" (Scheer, 2017, p. 105). Furthermore, scientific knowledge and its application to a specific context can also provide legitimacy to political action.

A common bridge between policy-makers and scientific knowledge are intermediary actors (Cole et al., 2021). The policy-making process itself is often very complex, with varying stakeholders involved. Policy-makers therefore tend to avoid adding further complexity or uncertainty inherent to scientific knowledge (Brugnach et al., 2007, p. 1085). Uncertainty itself, however, is unavoidable when generating knowledge aimed at informing policy, even more so when concerned with social agents and complex social systems.

Ahrweiler (2017) argues that even in highly-controlled and deterministic scenarios, prediction and therefore planning are sometimes impossible. If we add to this the complexities of social behaviour, it may seem futile to do any kind of planning at all (Ahrweiler, 2017). We are thus tasked to find alternative approaches for scientific knowledge to inform policy-making processes. Many have suggested that agent-based modelling (ABM) is such an approach (see e.g., Ahrweiler, 2017; Lempert, 2002). We subsequently introduce the methodology of ABM, before arguing how ABM can aid in informing decisions on policy and science policy in particular.

2.2. Agent-Based Modelling

Much quantitative research in the social sciences relies on models. Models serve many purposes, related to the representation of either theory, data, or both, to learning through manipulation, or strictly to accurate measurement (Morgan & Morrison, 1999). This is equally true in the case of ABM (Squazzoni, 2012, pp. 9–11). ABMs are defined as "a *computational* method that enables a researcher to create, analyze, and *experiment* with *models* composed of *agents* that interact within an *environment*." (Gilbert, 2008, p. 2)

A commonly shared goal of quantitative research in the social sciences is that of explanation (Abbott, 2004), e.g. explaining publication performance by reference to individual attributes such as gender, ethnicity, or academic age. With agent-based modelling, the approach is instead to generate patterns on the aggregate level (macro) based on bottom-up data that is generated by simulating individual agents (micro). This enables the researcher to test interventions and their repercussions for macro level phenomena by altering individual-level behaviour. However, it is important to note that such an approach might still fail to identify the true causes of macro regularities, since an ABM might fail to contain all *sufficient* conditions, by only including conditions and effects that are *necessary* to result in the expected empirical outcome. Claims about causal effects therefore need to be substantiated by "careful empirical inspection" (Squazzoni, 2012, p. 12), but also by taking into account all relevant assumptions made by the model (Murray et al., 2017).

Despite these caveats, which in fact apply to most types of scientific modelling, ABM offers an intriguing approach to the "problem" of the micro-macro link. The question of how individual action and the emergence of macro-level phenomena are to be linked has deep roots in the tradition of sociological theory (Alexander et al., 1987). In its essence, this problem amounts to two fundamental questions: (1) are aggregate outcomes the mere sum of individual actions (methodological individualism), or do emergent phenomena which transcend individual actions exist (norms and social facts, following the definitions of Durkheim)? And (2) how can individual action explicitly be linked to the emergence of macro-level phenomena? ABM offers a pragmatic approach to these questions, by explicitly linking the actions of individual agents (micro) to outcomes at the system level (macro). The model then allows investigation of both, how individual action affects collective outcomes, and how structural factors lead to individual outcomes (Bruch & Atwell, 2015; de Marchi & Page, 2014; Squazzoni, 2012).

Another important benefit of conducting ABM is its relative simplicity, compared to alternatives such as purely mathematical modelling. In analyses that combine agents with heterogeneous beliefs, multiple interacting mechanisms, and pre-existing or emerging network structures, mathematical models might be very difficult to develop (Watts & Gilbert, 2011). What is more, ABM also allows for the emergence and analysis of nonequilibrium processes, which is a common assumption in mathematical modelling (de Marchi & Page, 2014).

2.2.1. Method

At its heart, ABM is exactly what the term alludes to: modelling based on individual agents. With individual agents, we refer to any entity whose actions or behaviour we want to model. This can be anything from individuals to institutions, to countries or groups of countries. Individual agents are instantiated with certain attributes and behavioural rules. Depending on the goals of the model, an individual agent's attributes and behaviours can be assumed to be fixed or to be mutable. Behavioural rules can be simple (e.g., only depending on the configuration of the model at a specific time) or more complex, involving agents considering past or possible future actions. Besides the individual agents and their characteristics, an ABM can also incorporate further aspects such as network structures and geographical information. The final piece to incorporate into an ABM is usually to implement interactions between the agents (de Marchi & Page, 2014, pp. 7–8). Based on this model, simulations can be run, the output of which is then investigated and visualised.

2.2.2. Calibration and Validation

Employing ABM can have multiple goals, from efforts to understand certain mechanisms under restricted conditions, to modelling complex social situations (Bruch & Atwell, 2015). Increasingly, ABMs rely on vast amounts of empirical data, especially when the aim is to develop policy (Ahrweiler, 2017; Taghikhah et al., 2021). In these cases, empirical data is used to calibrate many aspects of the model, such as the individual agent's attributes and their behaviour. Such "high-fidelity ABMs" (de Marchi & Page, 2014) have been heavily used in research on the COVID-19 pandemic (Lorig et al., 2021; Squazzoni et al., 2020).

On the other end of the modelling pipeline, a common approach is to validate the outputs of an ABM by comparing them to known data/patterns in the real world. This aspect usually is an integral part of building the model, since if the aim of a given modelling exercise is to shed light on some aspects of reality, the model will only be of use if it can in principle reproduce key empirical characteristics of the situation/phenomenon under scrutiny. On the other hand, if the goal is solely to explore certain mechanisms and their conditions, a fully conceptual model without reference to further data will often still be sufficient.

2.3. ABM and policy making - benefits and challenges

As discussed above, policy-making usually has to deal with high levels of uncertainty and contingency. Social simulations provide a powerful way to address these issues. Although ABM does not offer certainty, inspecting the model and its output allows the researcher "to see the options and possibilities of what can happen and what cannot." (Ahrweiler, 2017, p. 398) Through such learning it is possible to investigate potential outcomes, pose "what-if"-questions and therefore shed light on counterfactual pathways. These benefits of ABM are especially important, given that social situations commonly exhibit complex dynamics, with feedback loops, self-reinforcing dynamics and unintended consequences, in cases where agents change their behaviour in response to interventions (Bruch & Atwell, 2015, p. 5).

While a large part of ABM for policy-making has focused on highly realistic models, backed by rich empirical data, also more stylized models can inform situations of decision making in meaningful ways. Bullock (2016) argues that models do not have to be empirically grounded and do not have to make precise forecasts to be useful for informing policy decisions. Rather, models (e.g., ABMs) help in understanding mechanisms that let one make more informed decisions about policy interventions.

Common strands of critique against the use of scientific simulations for policy making, which also apply to ABM, are concerned with questions of model assumptions, starting conditions, and the inaccessibility of model output (figures and numbers) for policy-makers (Scheer, 2017). To meet the challenges inherent to the modelling itself (in terms of assumptions and starting conditions), ABMs targeted at informing policy-making processes in the area of science policy, such as the SKIN model (Ahrweiler, 2017; Ahrweiler et al., 2015), integrate rich data and aim at reproducing empirical phenomena in a baseline scenario. If this is successful, then, following the logic of mechanistic explanation which underpins the use of ABM, changes to the simulated system are expected to reflect true changes as if interventions had taken place in the real world.

Finally, the potential inaccessibility of model output and derived conclusions again highlights the importance of intermediary actors who bridge the gap between scientific knowledge and the needs and demands of policy-makers (Cole et al., 2021).

2.4. ABM for Science Policy

While ABM has been used to inform policy decisions in a wide range of settings, a particular strand of research has focused on ABM for science policy. Salient examples include the SKIN model (Ahrweiler, 2017), models on how to better distribute research funding (Bollen et al., 2017), models on the transition to Open Access (Bernius et al., 2013), and models to investigate the reason for and spread of problematic research practices (Higginson & Munafò, 2016; Smaldino & McElreath, 2016), which we describe in detail below.

Over the course of multiple research projects, Ahrweiler and colleagues (Ahrweiler, 2017; Ahrweiler et al., 2015) developed an agent-based modelling platform named "SKIN", which is concerned with "**S**imulating

Knowledge Dynamics in **I**nnovation **N**etworks" (Ahrweiler et al., 2015). The model comprises agents (research institutions, large diversified firms, small and medium enterprises) that seek to apply for funding by the European Commission's seventh framework programme (FP7). Agents form consortia and deliver research outputs based on their specific capabilities, which evolve through collaboration. Ahrweiler et al. (Ahrweiler et al., 2015) show that the resulting network structures closely match observed structures, which enables them to model potential interventions to the funding system. The outcomes of potential interventions were then shared and discussed with policy-makers from the European Commission, and subsequently influenced policy decisions on the subsequent H2020 funding scheme.

Higginson and Munafò (2016) investigated how current incentive structures in science lead to a proliferation of underpowered studies. In their model, researchers are tasked to decide to which extent to spend their fixed resources on exploratory and confirmatory studies, as well as which sample sizes to target. Higginson and Munafò find that, given current emphasis placed on novel findings over confirmatory studies, or simply having many publications, it is optimal for researchers to conduct many exploratory studies with low sample sizes and therefore low statistical power, which leads to erroneous conclusions for about 50% of all studies. The authors recommend to assign lower weight to novel findings and key publications in evaluating researchers for appointment and promotion, and to strengthen statistical requirements in editorial decisions on publishing or rejecting manuscripts, by requiring higher sample sizes and lower α-levels for statistical tests (Higginson & Munafò, 2016, p. 10).

In a similar vein, Smaldino and McElreath (2016) used a dynamic model of science to show how the high value placed on the total number of publications produced by individual researchers and research groups leads to the spread of poor methods. They assumed no bad intent on the side of researchers, but simply operated on common evolutionary dynamics: Research groups with higher output receive higher rewards, and are therefore more likely to generate “offspring”, as in former students creating new labs, or by influencing other scholars in the same field. Their model shows that this mechanism by itself leads to the spread of poor research methods, while efforts to replicate findings and thus “expose” poor methods are able to slow, but never halt or reverse the dynamic. A difficulty encountered in Smaldino’s and McElreath’s use of ABM was to show the “precise relationship” between some quantities under study, “because all these parameters are entangled” (Smaldino & McElreath, 2016, p. 13)**.** They opted to create a simplified version of their model, omitting any evolutionary dynamic and fixing various parameters at certain levels to isolate specific effects.

2.5. Data Sharing - State of Play

Policies for data sharing now exist at multiple levels within the scientific system. Funders increasingly encourage and sometimes mandate the sharing of primary research data. The European Commission is pushing for FAIR (Findable, Accessible, Interoperable, Reusable) and Open Data in Horizon Europe, with explicit calls for research data management in compliance with the FAIR principles and to ensure access to research data following the principle of “as open as possible and as closed as necessary”[4]. At the level of journals and publishers, policies for data sharing are also becoming more common, albeit starting out from low levels (Vasilevsky et al., 2017).

[4] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

To date, large-scale assessments of current levels of data sharing across disciplines are scarce. Serghiou et al. (2021) conducted a landmark investigation into various aspects of transparency related to scientific publishing, including data sharing. Across 2.75 million Open Access articles from PubMedCentral, they found an increase in the rate of articles sharing data, with an estimated 15% of articles published in 2020 sharing research data.

Statements of data availability alone, however, will not lead to increased reuse of research data. A key concern in this regard is the quality of the shared data. Quality here mainly refers to adherence to the FAIR principles (Wilkinson et al., 2016). The crucial question in this regard is whether data is merely shared to fulfil mandates, without much consideration as to how the data might be re-used by someone else, or whether sufficient provisions are taken to enable others to appraise and potentially re-use the shared data with little effort. Hardwicke et al. (2018) investigated the effect of a journal introducing a data sharing policy and found, as a consequence of the introduced policy, a substantial increase in the number of articles that shared data (from 25% pre-policy to 78% post-policy), and in particular, an increase in the share of articles where the shared data was “accessible, complete and understandable” (22% pre-policy, 62% post-policy).

Similarly, Tedersoo et al. (2021) investigated data sharing across all articles from *Nature* and *Science* from 2000-2019. Both journals have comparatively stringent policies on data sharing, and the authors subsequently found high average levels of data availability (before contacting authors: 54.2% for 2000-2009, 71.8% for 2010-2009), with substantial variation between fields. On contacting the authors of initially unavailable data (e.g., from articles stating data was “available upon request”), data availability increased substantially, to 70% (2000-2009) and 83% (2010-2019). However, only about 40% of data that was stated to be available upon request was eventually obtained. Tedersoo et al. (2021) therefore conclude that the option to provide data only upon request is ineffective and should be abandoned by journals. To increase rates of data sharing, they recommend that data sharing be associated with “real benefits such as recognition, or bonus points in grant and job applications” (Tedersoo et al., 2021)**.** Other proposed incentives to increase rates of data sharing among various disciplines focus on aspects like elevating credit gained from data sharing, either through recognizing data sharing via authorship, awarding badges to papers sharing data, tracking data sharing via specific metrics, as well as introducing mandates for sharing data at journal or funder levels (Woods & Pinfield, 2021)**.**

Despite various benefits to data sharing at the individual and the system level, making research data FAIR is evidently still not common in most disciplines. Among numerous factors contributing to this trend, the resources and effort required to adequately share data are a major contributor. Although data sharing is often pitted as an individual’s decision, it involves multiple stakeholders at various levels, and must therefore be understood as a collective undertaking (Borgman & Bourne, 2021). However, levels of institutional support differ and might lead to differential rates of research data sharing (Ross-Hellauer et al., 2021)). What is more, not only the sharing of data, but also its reuse is linked to institutional support, through training, funding and computational resources necessary for making effective reuse of data (Johnson, 2018). Policies to increase sharing and reuse of data therefore might disproportionately benefit already advantaged institutions and individuals (Ross-Hellauer et al., 2021).

2.6. Summary

Over recent years, various proposals have been made to increase the extent of researchers sharing the data their findings rely on. Some interventions, e.g. journals requiring data availability statements, have been shown to increase rates of data sharing (Hardwicke et al., 2018), while for others, e.g. awarding badges for sharing data alongside publications, results have remained inconclusive (Woods & Pinfield, 2021). As with any intervention in non-deterministic systems, anticipating outcomes and their repercussions is complex. However, given the stakes involved in how policies shape the modes of scholarly communication, suggested interventions aimed at increasing the sharing of data should be probed for their efficacy and any unintended consequences. As we have discussed above, ABM offers a coherent framework to conduct such modelling, taking into account complex logics of costs, benefits, and individual propensities for changing practices.

With the DASH model, we leverage the potential of ABM to investigate how plausible scenarios of incentivising data sharing could encourage its uptake, given varying scenarios of costs associated with the sharing of data, as well as different strategies employed by individual research groups to decide on whether or not to share data. The model will serve as a starting ground, enabling further evaluation of policy interventions for their efficacy, and potential repercussions for equity in science.

3. Methods

The aim of the DASH model is to investigate the uptake of data sharing among a hypothetical scientific field. We model a basic publication process, where research groups produce publications based on their resources. Resources are assumed to have a constant baseline. Additional resources are available through grant funding. We first model a baseline scenario that does not include the sharing of data. With this baseline scenario, we model the process of cumulative advantage, where some groups are able to secure further resources based on existing ones. To this process, we then add a process for groups to share data, as well as mechanisms for them to revise their behaviour. Equipped with this model, we then analyse the impact of varying costs and incentives for the sharing of research data on its uptake. The full code for the model and the analysis is available at https://doi.org/10.5281/zenodo.5913719.

3.1. Model assumptions

The unit of analysis for the model are research groups.[5] In the era of team science (Latour & Woolgar, 1986; Lee et al., 2015; Price, 1963), research groups form around principal investigators (PIs). If successful, these groups generate further funding through grants, thus expanding over time. The expansion in resources and personnel allows for an increase in publication output, and in many cases in subsequent impact.

We model agents from one hypothetical field, with a fixed number of research groups. To represent time, each round is defined as lasting 6 months. Besides research groups, we assume a single funding body that is able to award 14 research grants per round.

Parameter name	Parameter description	Default value
n-groups	Number of research groups	100
history-length	Length of publication history considered for funding decisions (rounds)	6
resources	Baseline resources that all groups keep irrespective of their publication success	0.3
n-funders	Number of funders	1
grants-per-funder	The number of grants each funder awards per round	14
randomness parameter *c*	A parameter used to avoid complete path-dependency when allocating funding	0.4
rdm-cost	Cost factor which reduces available resources when sharing data	0.05
data-sharing	Boolean value representing the current choice of a given research group whether or not to share data	Initially TRUE for 50% of groups
agent-orientation	Whether groups are myopic, long-term oriented or a mixture when comparing current to past success	all-myopic, all-long-term
openness-for-change	Probability of a group changing their sharing decision if the best peer group is more successful (applied in the model on social learning)	0.2

[5] We choose research groups, because it would not be plausible to observe effects of cumulative advantage over long time periods for single researchers.

Table 1: Model parameters.

3.2. Agent attributes

3.2.1. Research groups

Research groups have one core attribute, their base funding rate which is modelled as constant. This reflects the assumption that research groups form around a single researcher (e.g., a professor) who is tenured and has sufficient resources to conduct their research. The base funding rate (0.3) allows for publication activity irrespective of any further funds that might be secured through the grant allocation process described below.

3.3. Publication process

In each round, research groups will try to produce publications. The number of publications generated by any group is drawn from a Poisson distribution. The parameter λ indicates the average number of events in the given time interval, and varies according to the following scenarios:

1. If the group currently is not sharing data[6], λ equals the base funding rate (0.3) plus the number of currently available grants (λ = 0.3 * n-grants).
2. If the group currently is sharing data, λ is calculated as above, but immediately (before producing publications) reduced by the factor of rdm-cost (0.05) (λ' = λ * (1 - rdm-cost)). This cost factor models the direct cost of undertaking state of the art research data management practices, and is estimated to be about 5% (Mons, 2020).[7]

It can be seen that for groups with no grants which do not share data, the expected mean publication rate is 0.3 publications per round. This value was established during efforts to calibrate the model against real-world data (see section 4.1). Each additional grant increases λ and therefore the expected mean publication rate by 1, since the mean of a random variable that follows a Poisson distribution is equal to λ.

3.4. Mechanism for resource allocation - grant funding

Each round, research groups have the chance to receive grant funding. A research grant lasts for 6 rounds (3 years), and provides additional resources that enable further publications. Each round, a fixed number of grants is available from each funder (the default assumes 100 research groups, one funder, and 14 available grants per funder per round). Groups are ranked by the total strength of their grant proposal, and the top groups are awarded a grant until no grant is left.

3.4.1. Baseline implementation

In the default model, grants are awarded based on recent publication performance and a random parameter *c*. Parameter *c* reflects the fact that funding decisions are not solely based on publication history, but also on proposal quality and other potential factors. Furthermore, allocation of funds in academia at large is more complex than our implementation of the funding mechanism. The parameter *c* allows us to capture this discrepancy, and to arrive at similar distributions of the number of published papers across groups, as can be found in the literature.

[6] See section 3.5 below on how groups decide whether or not to share data.

[7] Costs for research data management can be expected to be variable across research fields, but we base our 5% level on Mons' (2020) calculation that on average, 5% of overall research costs are required for adequate data stewardship and ensuring data reusability."

To calculate the strength of grant proposals, we first denote publication performance as p, which is the total number of publications over the publication history (h_p). The length of h_p can be set to any value between 1 and 20 years, and is set to 3 years by default.

The parameter c is drawn from a uniform distribution on [0, 1]. Publication performance is scaled to the range [0,1] by dividing it through the maximum publication performance of all groups in a given round.

The weight given to publication performance and c can be varied, ranging from only taking publication performance into account, to solely awarding grants based on chance. By default, we assume the total strength of a proposal to be determined to 40% based on c, and 60% based on publication history ($c_w = 0.4$). The value for c_w was developed during model calibration (see section 4.1).

Total strength of a proposal is thus calculated as

$$total\_strength = c_w * c + (1 - c_w) * p$$

3.4.2. Incentivising data sharing

To investigate the impact of incentivising data sharing through funding decisions, we slightly modify the calculation of proposal strength from above. Instead of solely relying on publication history as a proxy of the success of research groups, funders also take into account the recent history of data sharing of the applicants. This shift can take the form of awarding grants solely based on data sharing history (and c), or a mix between data sharing history and publication history.

Formally, the strength of proposals in this case is calculated as

$$total\_strength = c_w * c + (1 - c_w) * r$$

where

$$r = p_w * p + (1 - p_w) * d$$

where d denotes the number of data sharing successes, calculated as the sum of datasets published over the history length, and $p_w \in [0, 1]$ denotes the weight given to publications compared to datasets.

3.5. Data sharing

In our model, the decision whether to share or not to share data is made by research groups. To allow for some baseline differential dynamics of resource accumulation to emerge in the system before the introduction of data sharing strategies, groups cannot share data for the first 99 rounds. At model initialisation, some groups (50%) are assigned the trait of sharing data. From round 100, these groups start to share data. They are then open to revise their decision whether or not to share based on two distinct mechanisms: individual learning, where they compare their current success with their past success, and social learning, where they learn from the success of others. Both mechanisms assume different time perspectives: myopic groups revise their decision every round, while more long-term oriented groups revise every 5 rounds. The differences in time perspective represent varying strategic positions of research groups: newer groups might be more focused on generating many publications in a short span of time than more established groups, which might take more long-term oriented decisions.

3.5.1. Individual learning

Under the assumption of individual learning, groups compare their success in terms of publications with their prior publication success. Myopic groups do this every round, while more long-term oriented groups compare their outcome of the last 5 rounds with the outcome of the prior 5 rounds.

The rule for this comparison is as follows:

```
if:   success-of-this-period * 1.5 < success-of-last-period → change behaviour
else: keep behaviour
```

The decision to include the multiplicative factor in the decision was governed by the presence of substantial variation in the publication process. Because current resources (from base funding and grants) are not 100% correlated with publication outcomes, some groups would change their behaviour, although the change in their success was simply due to the stochastic process of drawing publications. In this sense, the multiplicative factor makes groups more conservative in changing their behaviour less often.

3.5.2. Social learning

For the case of social learning, groups assess the success of others and copy the data sharing decisions of the most successful group. Each time groups update their decision, they first randomly draw 5 other research groups to compare with. The order in which groups revise their decision is random. Second, they rank the other groups according to their publication success, and inspect the current decision of the most successful group. If the successful peer group has a different state of data sharing, they copy the behaviour with a probability of 20%. This reflects the fact that research groups might exhibit a certain inertia in updating their decision.

The time-orientation of agents (myopia vs. long-term orientation) here only affects how frequently agents update their decisions. They still compare their peers according to the current publication success (as measured by the current number of publications in the current round). However, since agent decisions “stick” for multiple rounds, interventions have enough time to exhibit an effect that can be picked up by the agents.

4. Results

4.1. Baseline scenario calibration

The goal of the DASH model is to unveil important characteristics of data sharing environments: how data sharing behaviour is influenced by costs associated with RDM, how incentives at funder level can counteract these tendencies, and how such interventions might reshape structures of resourcing among research groups. To reach this goal, the model does not have to be a perfect representation of reality. However, to ensure the model accurately mirrors the key features of the environment targeted, we calibrate the model using real-world data. Our target in this regard was the distribution of publications across research groups, (a) because this is the main baseline outcome of our model, and (b) because publication data is generally available to use as a reference.

The number of publications of individuals tends to follow a power-law distribution, often referred to as Lotka's law (Lotka, 1926): many individuals have few publications, and few individuals have many publications. While this distribution has been shown for individual researchers, similar distributions can be observed on aggregated levels. Analysing data on more than 1,000 universities from the Leiden Ranking[8] (Van Eck, 2021), we find the characteristic distribution also for large institutions (Figure 1). Computing the Gini index for the number of publications across universities, we obtain values from 0.47 to 0.53. Two key observations can be made: first, the level of inequality in publication counts across universities is declining every year. Second, the share of institutions with very few publications is declining as well.

[8] The data from the Leiden Ranking are provided in 4-year segments.

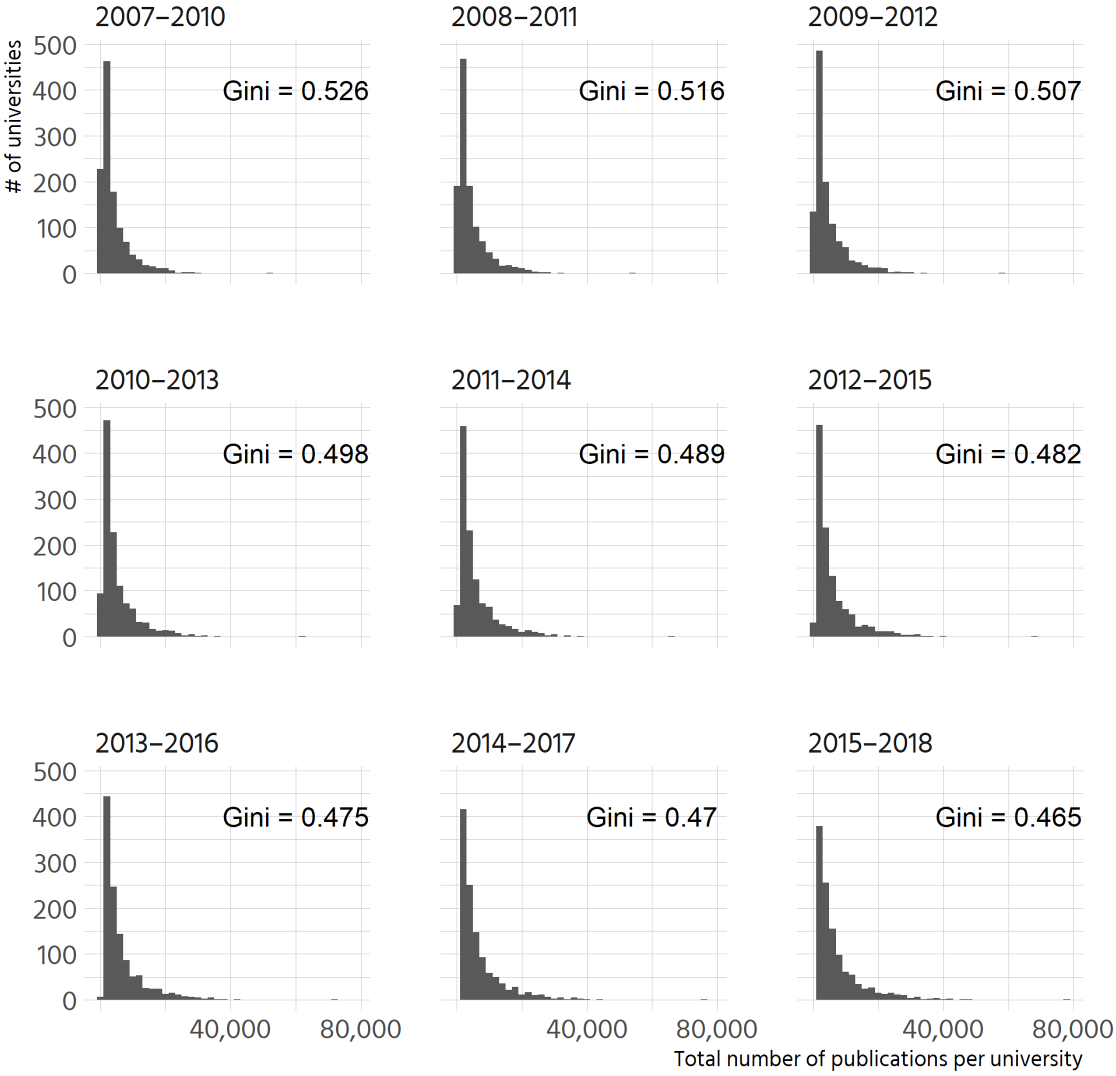

Figure 1: Distribution of publications per university (2007-2018). Data from the Leiden Ranking.

Data on publications of research groups is not easily obtainable, first and foremost since the definition of research groups involves non-trivial methodological questions on how to define research groups. Franceschini and Maisano (2011) collected data on 33 Italian universities, defining research groups pragmatically as all researchers of a given university working in the field of "Production Technology and Manufacturing Systems". This approach likely overestimates the size of research groups, since larger departments might comprise multiple research groups, while small departments might also be conceptualised as single researchers working independently, rather than a unified research group. Nevertheless, the data are informative and relevant for our purposes. Figure 2 displays the distribution of publications of the 33 research groups from the paper by Franceschini and Maisano. Similar to the data from the Leiden Ranking, the Gini index for the distribution of publications across the 33 research groups is 0.56.

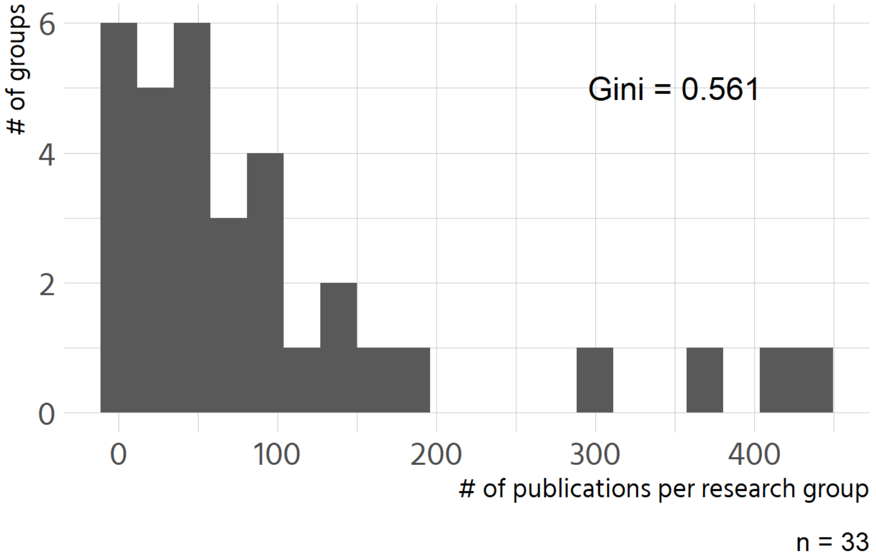

Figure 2: Distribution of publications of research groups as reported in Franceschini and Maisano (2011).

The baseline model presented above allows for the adjustment of multiple parameters: the number of research groups, the number of funders, the number of grants per funder, the length of the publication history considered by the funders, as well as the role of other factors, subsumed in the random parameter *c*. Our aim was to reproduce the distributions of publications presented above. To this end, we fixed certain parameters and tuned the remaining ones to obtain similar results as in the reference data.

We assume 100 research groups, mainly to constrain computational demands to manageable levels. Further, we assume the length of the publication history considered by funders to be 3 years. This time span is a reasonable assumption based on the fact that funders tend to take into account recent publication history when assessing grant proposals (Hicks et al. 2015). Subsequently, we adapted the number of grants per funder and the influence of random factors on the model to get close to our target distributions.

Regarding the number of grants per funder, we assessed values between 1 and 20 over a time span of 500 rounds. Fixing the parameter *c* (randomness) at 0.5, having only one available grant per round leads to a scenario of high path-dependence, with a single research group monopolising most research grants over time (Gini index for the distribution of grants: ~1). Other groups do occasionally receive grants, but are not able to build on them. Due to the presence of base funding, the inequality within the distribution of publications is much lower (Gini = ~0.18). On the other hand, assuming 20 available grants per round, the distribution of grants is more equal (Gini = ~0.65), while the distribution of publications exhibits similar inequality to the scenario with only one grant per funder (Gini of publications = ~0.2, with a downward tendency).

Considering the model's sensitivity to the parameter *c*, we observed values from 0 to 1 (Figure 3). High levels of randomness lead to overall low inequality in the system, with a fairly normal distribution of publications across research groups. Low levels of randomness lead to higher levels of inequality, with a bimodal distribution in publications: some groups are able to receive funding early on and continue to receive funding, while all the others are not able to obtain grant funds. In tuning the model parameters, we sought to avoid this bimodal distribution by choosing values for *c* close to 0.4.

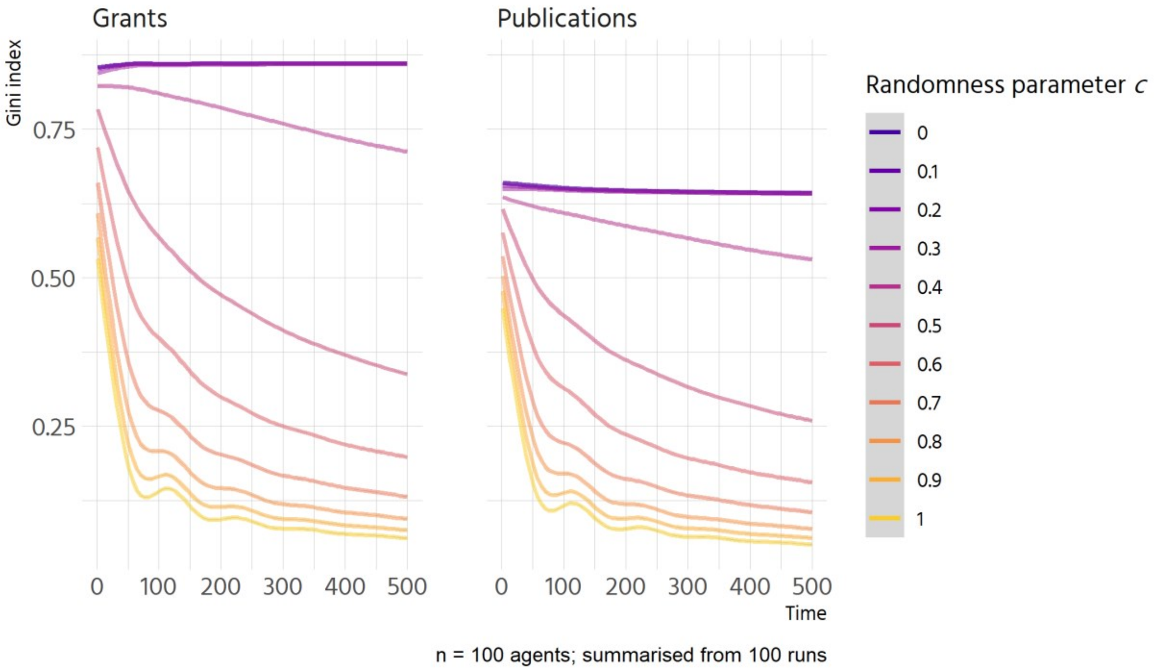

Figure 3: Gini indices for distributions of grants and publications. Here we consider the full range of values for the randomness parameter c.

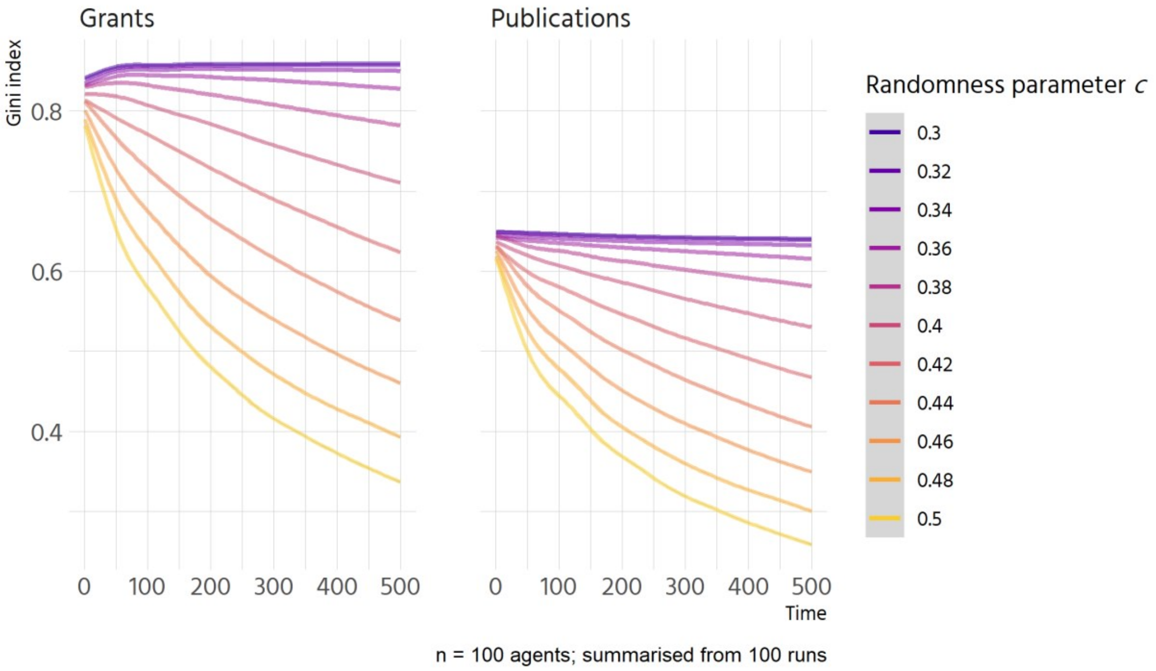

Figure 4: Gini indices for distributions of grants and publications. We consider a reduced range of values for the randomness parameter c.

For the final model, we settled on 14 available grants per round and a randomness parameter *c* of 0.4. This choice was based on the Gini index for the run with *c* = 0.4 covering a significant share of the observed Ginis from the Leiden ranking data and the data retrieved from Franceschini and Maisano (2011).

One potential way to make the model more realistic, and with that to reduce the parameter *c,* would be to introduce grants of different sizes (drawn randomly from an applicable distribution). However, this would add further complexity to the model. For this reason, we opted to subsume a range of factors in parameter *c*, which we do not investigate further. The distribution of publications and their respective Gini indices of 100 simulation runs for values of *c* between 0.36 and 0.44 are shown in Figure 5.

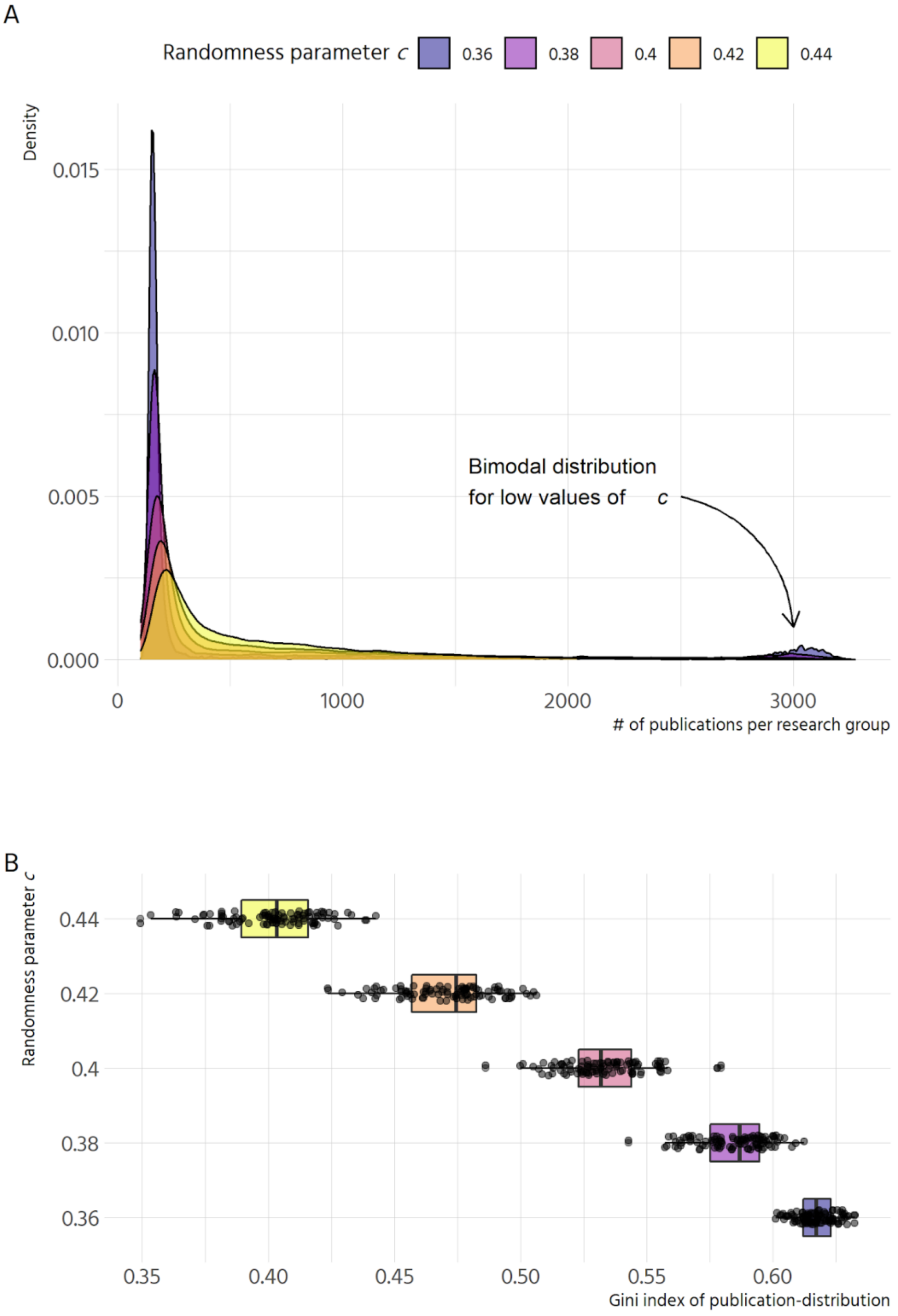

Figure 5: (A) Distribution of publications after 500 steps (aggregated from 100 runs). Values on the y-axis are based on a kernel density estimate, with a gaussian kernel. (B) Gini coefficients for the distribution of publications (after 500 steps).

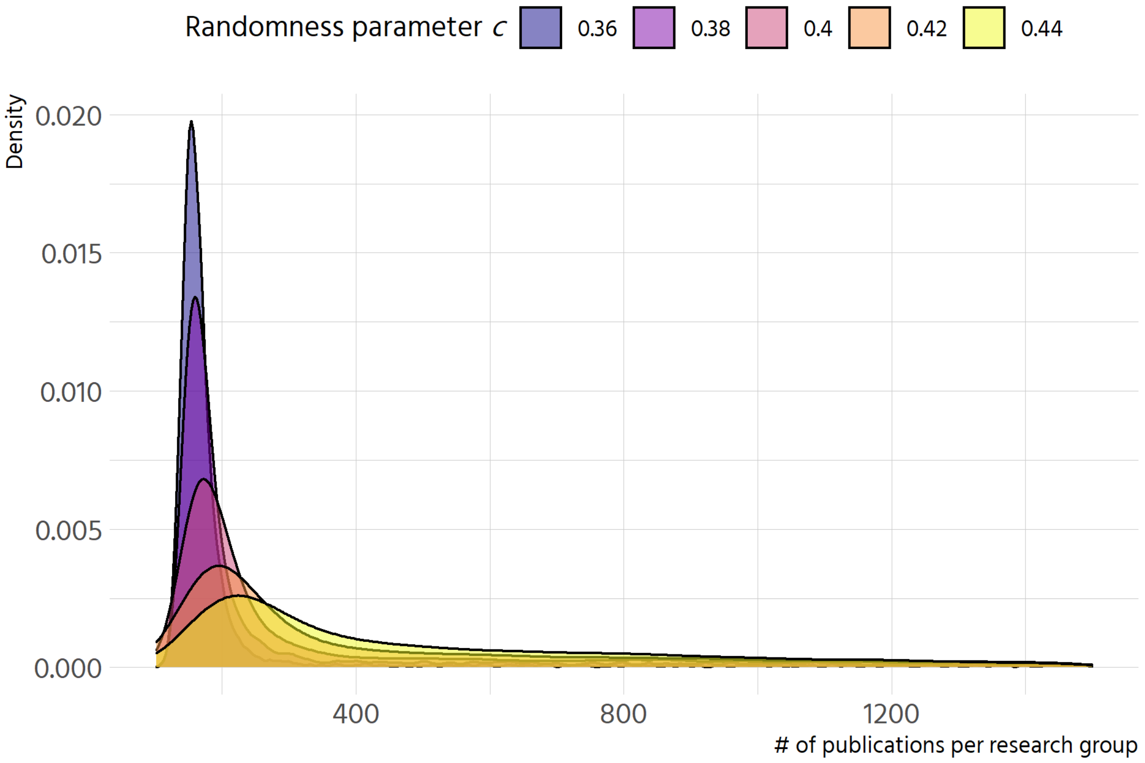

Figure 6: Detailed view of the publication distribution of 100 runs. X-axis truncated to 1500 publications.

4.2. Data sharing - Individual learning

Below we analyse the results of the mechanism of individual learning. We assess the following scenarios: (a) all agents are myopic (assessing only the last round), (b) all agents are long-term-oriented (comparing periods of 5 rounds), (c) a mix (comparing from one to five rounds). Furthermore, we assess the combination of the following policy scenarios: No incentives to data sharing from funders, and considering data sharing for 15% and 50% (compared to 85% and 50% publication history). Lastly, we assess three different cost factors for data sharing: no costs, 5%, and 10% costs inherent with data sharing.

4.2.1. Diffusion of sharing behaviour

Assessing to which extent agents revise their decision and adapt to incentives, we first investigate the share of research groups that choose to share data at any given time (Figure 7). Our results suggest agents only weakly adapt to cost factors and incentive structures under the given mechanism of individually comparing current to past performance. While myopic agents seem to make their decisions at random (given that the share of groups opting to share data always stays at 50%), long-term oriented groups adapt to incentives and costs to some extent. In a scenario with incentives for, but no costs of, data-sharing, a majority of groups opts to share data. Similarly, under no incentives and high costs (bottom right), fewer long-term oriented agents opt to share data. Both outcomes are as expected given the model specification, however the extent of the effects is much lower than initially anticipated.

A potential reason for this outcome is the contingency inherent in publication outcomes: with no grants, a research group is only expected to produce 3 publications in 10 rounds on average. Myopic agents (comparing round per round) therefore are unable to find more successful strategies, and even long-term oriented groups (considering 5 rounds) fare only slightly better. A potential solution might be to increase the timescales for comparison even further. However, this might also inhibit learning by itself, since agents get only few chances to update their decision over the course of the simulated time-window.

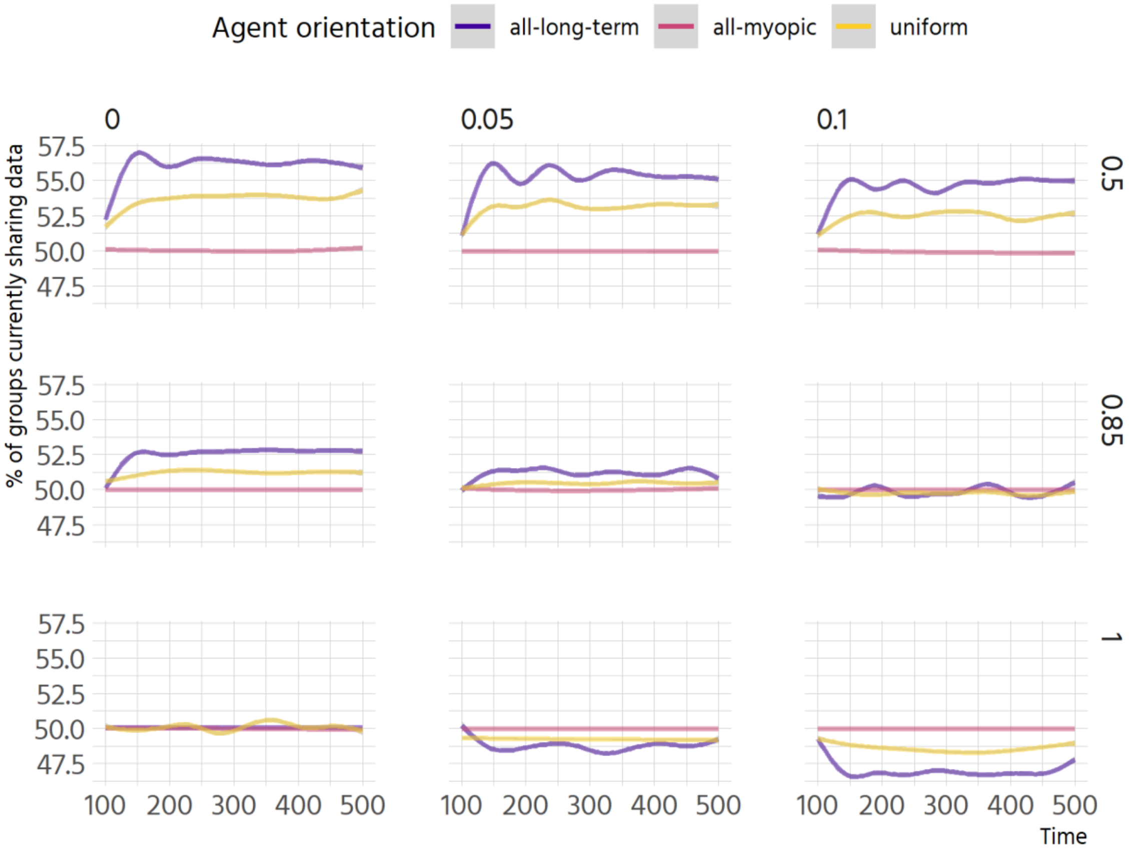

Figure 7: Number of agents sharing data. The columns (0, 0.05, 0.1, 0.2) refer to the cost factor inherent with sharing data, while the rows (1, 0.9, 0.8) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.2.2. Extent of data sharing

Following from the previous observation, the extent of available datasets is to some extent governed by research groups revising their decisions, but also determined by which groups on average tend to get rewarded, and which levels of sharing are possible given a certain cost factor. Figure 8 displays the number of available datasets per round (datasets expire after 10 rounds) by the same quantities as the above figure. Considering the cost factor of data sharing first, we observe lower rates of data sharing with higher cost factors, across all time-orientations. This follows directly from our model formulation, since the cost of data sharing effectively reduces the chances to produce publications, because some effort is spent on preparing and documenting data rather than preparing publications. However, as data is only shared when a publication is produced, fewer publications also mean fewer shared datasets.

The incentive structure (whether data sharing is considered in the granting mechanism) does affect which groups tend to share data to some extent (as discussed above), but it also affects which groups fare better.

Since the overall level of data sharing is higher when data sharing is incentivised through funding, this suggests that under the presence of incentives, groups that share data are more successful overall, thus leading to higher overall rates of data sharing (since these same groups receive more funding which enables further research and data sharing).

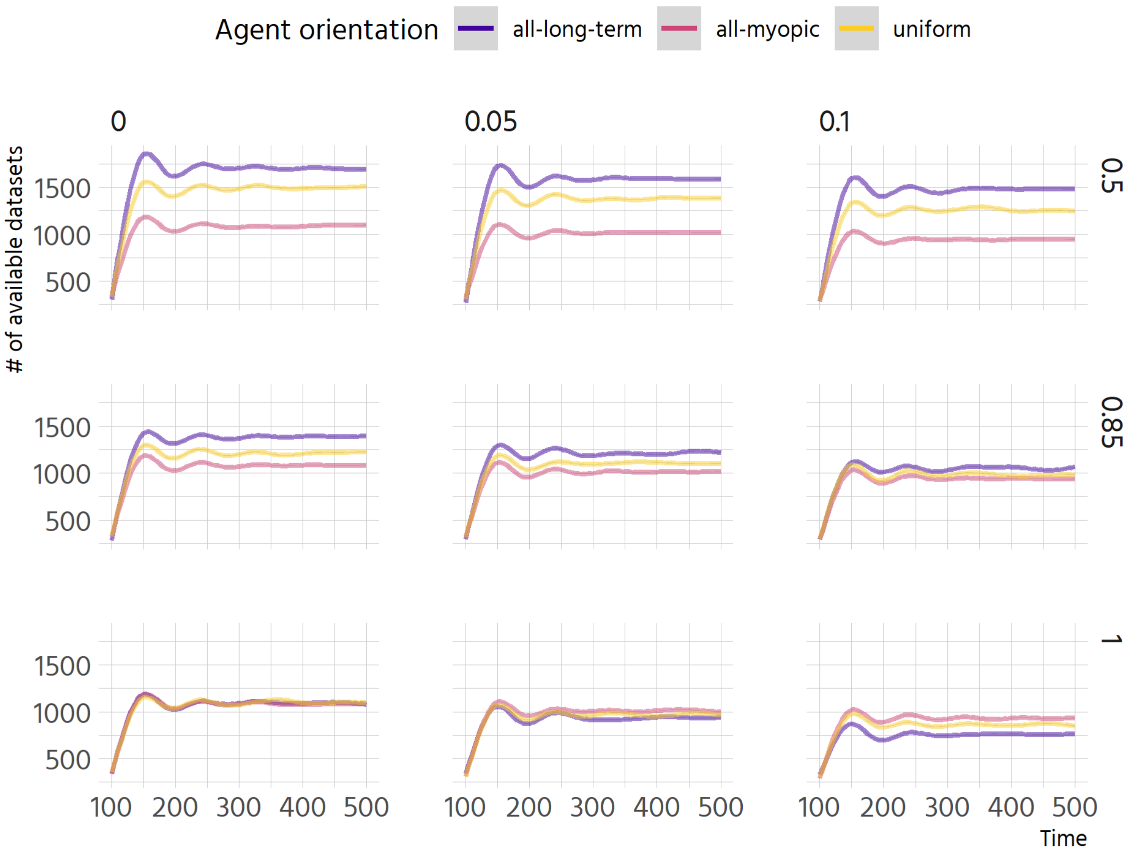

Figure 8: Number of available datasets. The columns (0, 0.05, 0.1, 0.2) refer to the cost factor inherent with sharing data, while the rows (1, 0.9, 0.8) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.2.3. Extent of publication activity

Considering the extent of publication activity of groups next, firstly, we observed no substantial difference in the trajectory of the number of total publications per our experimental conditions. However, there are some differences regarding the overall level of publication activity which accumulate to larger differences at the end of the simulation. Figure 9 thus displays the total number of publications at the end of the simulation, given our input parameters.

Similar to the above results, the agent's time orientation (myopic vs. long-term vs. mix) does make a difference in terms of how many publications are produced only in a few cases. In line with the model specification, higher cost factors for data sharing lead to lower total rates of publication. This process is moderated by the level of incentive structures. When funders reward data sharing, they tend to award grants more frequently to groups that share data, leading to a higher level of data sharing (as observed above) but a lower level of publication.

Some differences regarding agents' time orientations emerge, however. Considering the bottom right of Figure 9, under no incentives and with high costs, long-term-oriented agents are adapting to the cost pressure, opting to share data in fewer cases, and thus leading to higher overall numbers of publications.

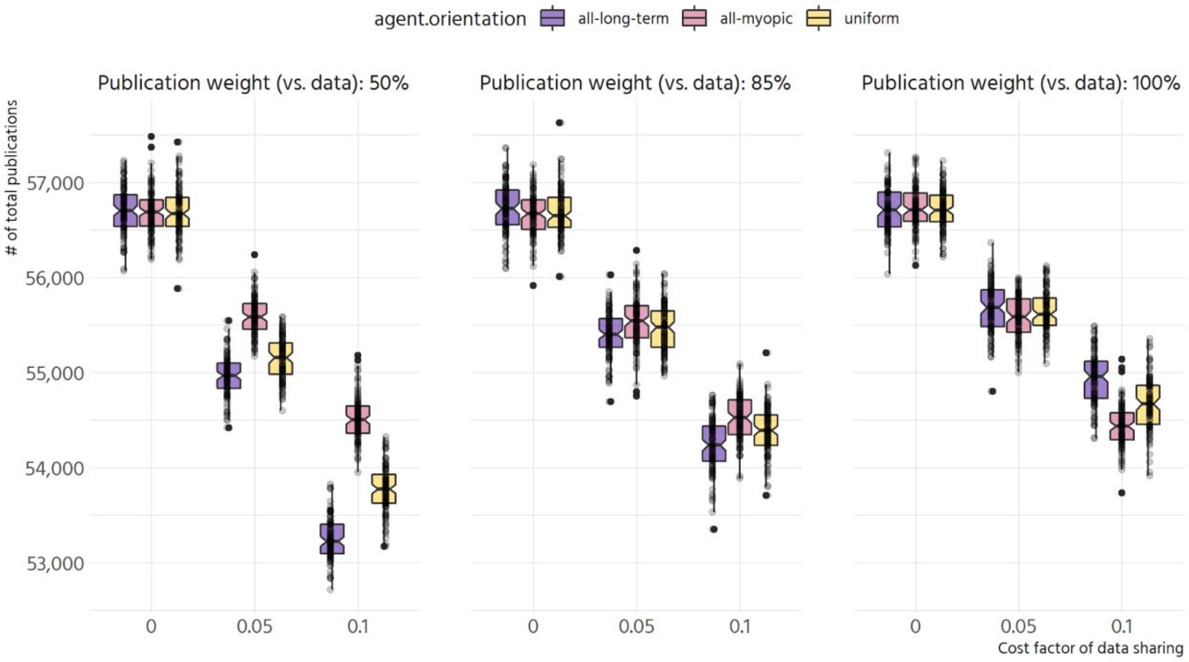

Figure 9: Total publications at simulation end. Results from 100 model runs per condition.

4.2.4. Equity of the system

We further investigate the overall equity of the system by considering the Gini index of datasets and publications produced, and grants held per each round. Figure 10 displays the Gini indices for simulation runs of long-term-oriented agents. The results for myopic agents are almost identical and not displayed here. Overall, the various input settings do not seem to have a substantial impact on the equity of the three distributions.

The level of concentration is higher for datasets than for publications, which mirrors the finding from above that not all groups choose to share data alongside their publications. Within the distribution of the number datasets produced per each round, higher cost factors for data sharing lead to a slightly higher concentration. There is an inverse relationship to incentives: higher incentives for data sharing are associated with lower discrepancies in the number of datasets published between varying degrees of RDM costs.

In analysing the findings from Figure 10, it must be noted that the observed concentration of datasets and publications produced *per each round* is substantially higher than the inequality in the overall number of publications and datasets which groups have produced. This can be explained by the contingency inherent in the model: while the distribution of datasets and publications *per round* is highly concentrated, these differences equal out over time, as most groups do not have an exceptional publication success each round.

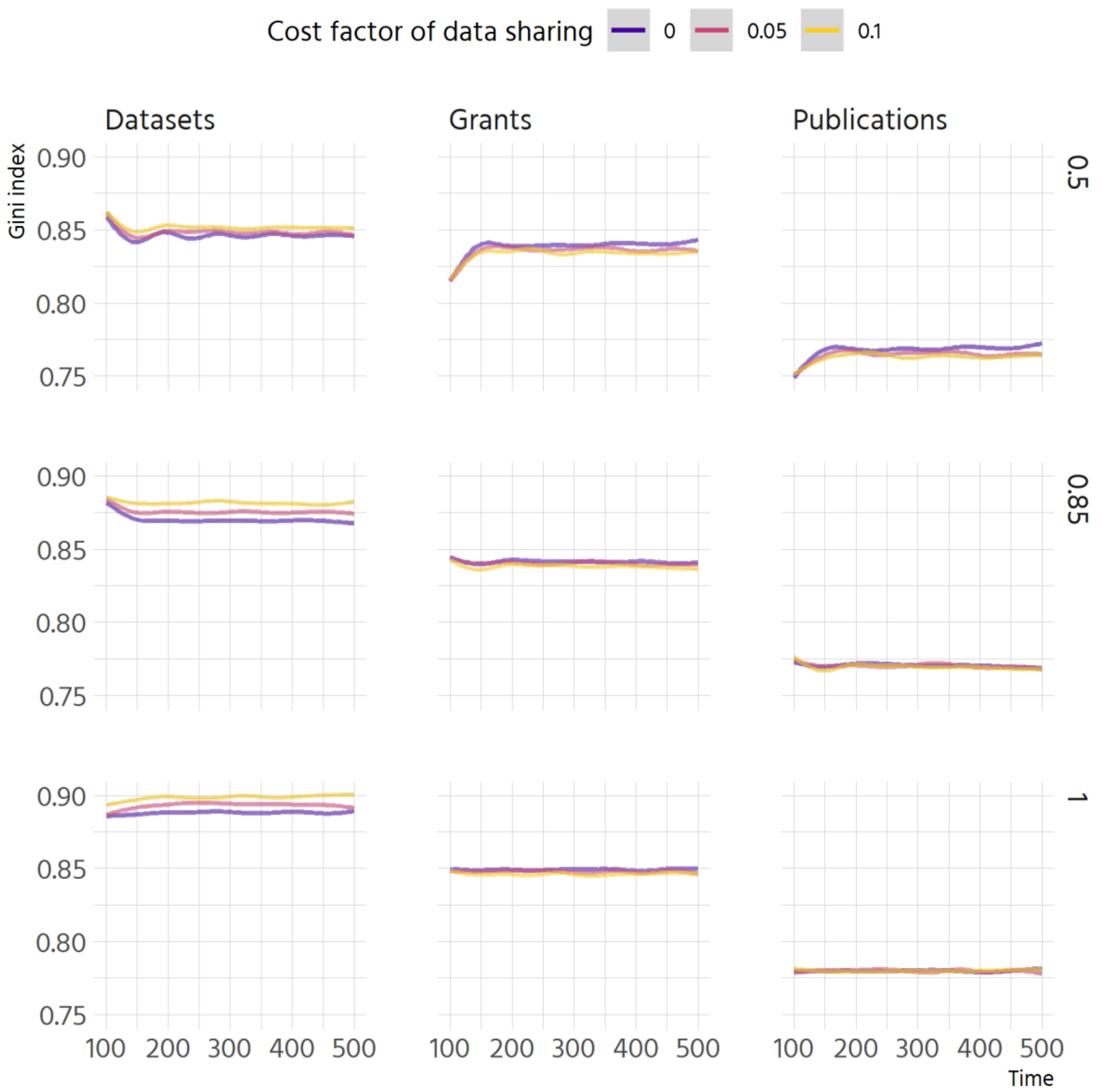

Figure 10: Gini indices of the distributions of datasets, grants and publications for individual level decisions. All agents with long term orientation. The rows (1, 0.9, 0.8) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.2.5. Behaviour of groups by publication success

During the initial phase of each simulation run, a characteristic setting emerges: some groups were initially successful in producing publications and subsequently managed to acquire research funding, while other groups did not. To investigate whether more successful groups make different choices than their less successful peers, we assign groups to quartiles according to the total number of publications produced so far, after the initial phase of the simulation (after 99 iterations). We then analyse how these groups react in terms of their decision to share data or not, in terms of the total number of datasets they produce, and in terms of their success in receiving further funding.

As described in section 4.2.1, groups only weakly adjust to incentives and costs under the individual learning mechanism. As Figure 11 shows, this adaptation mainly takes place in the most productive groups. However, the initially strong reaction to stop or start sharing quickly tapers off, and groups that were most successful until round 100 do not share data more or less frequently than other groups by the end of the simulation. This finding corroborates our previous interpretation, in that more successful groups are initially able to recognize better strategies due to their higher publication output. However, this initial advantage quickly dissipates due to incentives not having a large influence on the groups' performances in the short and medium term.

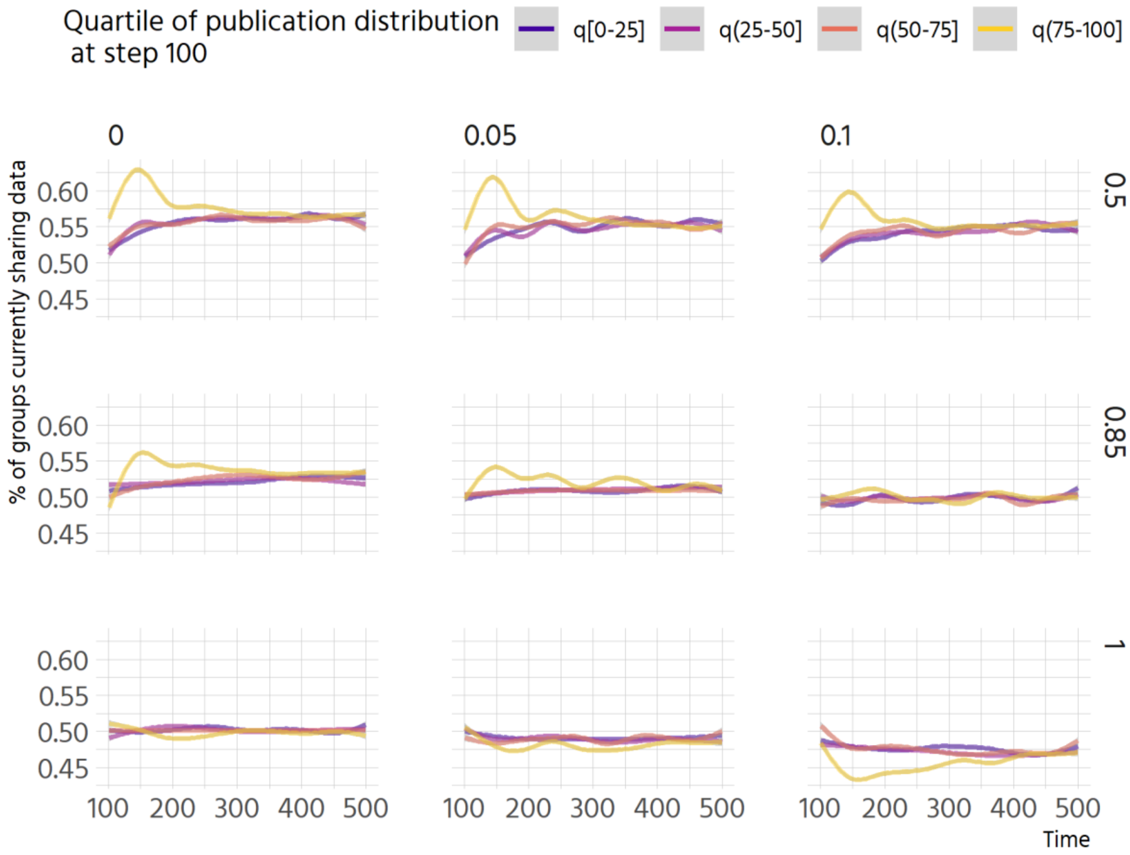

Figure 11: Proportion of groups sharing data by publication quartiles with individual learning. All long-term oriented agents. Aggregated from 100 runs.

Regarding the total number of datasets produced, we observe a clear divide between the most successful groups (q(75-100]) and all other groups (Figure 12). These differences cannot be explained by changed behaviour (since in many scenarios groups share at an equal rate), but are attributable to overall levels of productivity. Research groups that developed an initial advantage in their output of publications and subsequently acquired multiple grants continue to produce much more publications and thus also datasets than all other groups.

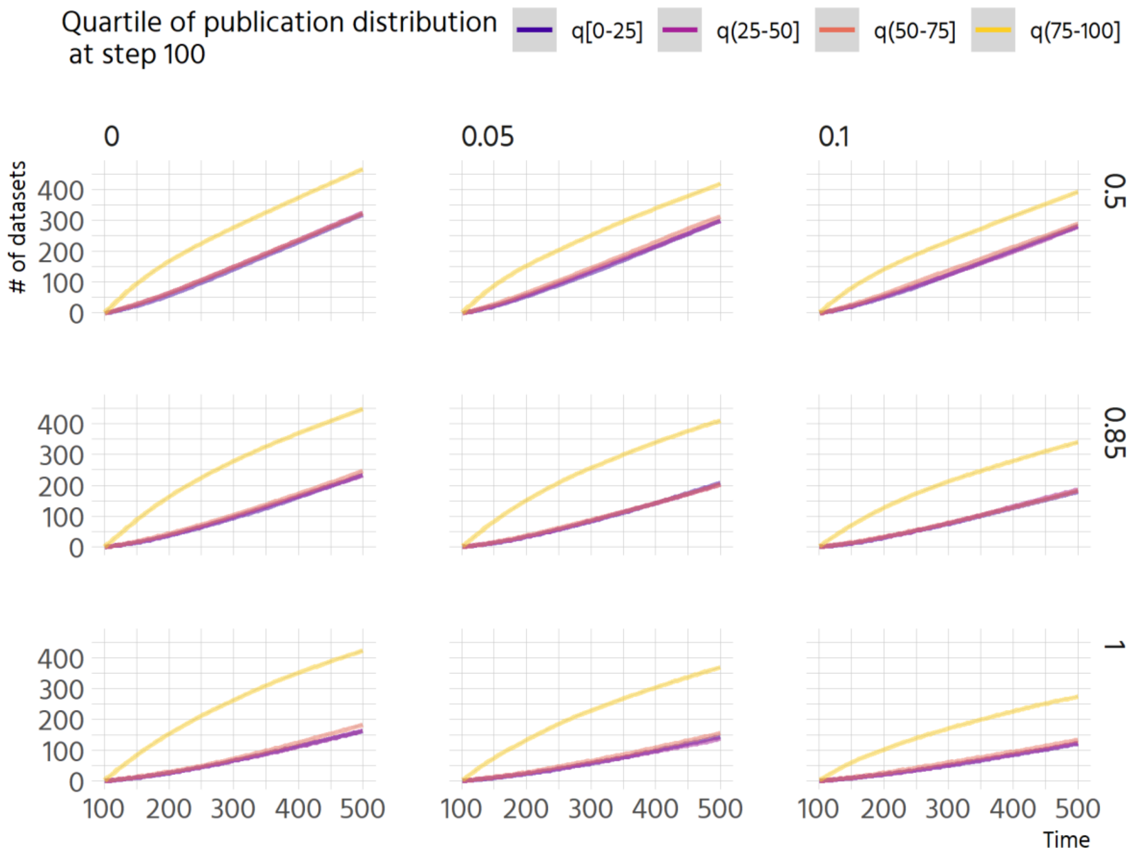

Figure 12: Mean number of total datasets per group with individual learning. All long-term oriented agents. Aggregated from 100 runs.

Considering the number of grants, we find small differences except for the best-performing quartile (Figure 13). Across all configurations, the average number of grants in the top group declines over time, which indicates that not all groups which are initially successful remain to be equally successful. The rate of this decline varies with the weight given to datasets in deciding on funding: with an equal weight given to publications and datasets, the decline in success of the top groups is much faster. This finding indicates that incentivising data sharing in funding decisions enables previously unsuccessful groups to receive funding more easily than in settings with low or no incentives for data sharing. This can be explained by the fact that incentivising data sharing reduces the competitive advantage some groups have developed by successfully publishing over a substantial number of rounds.

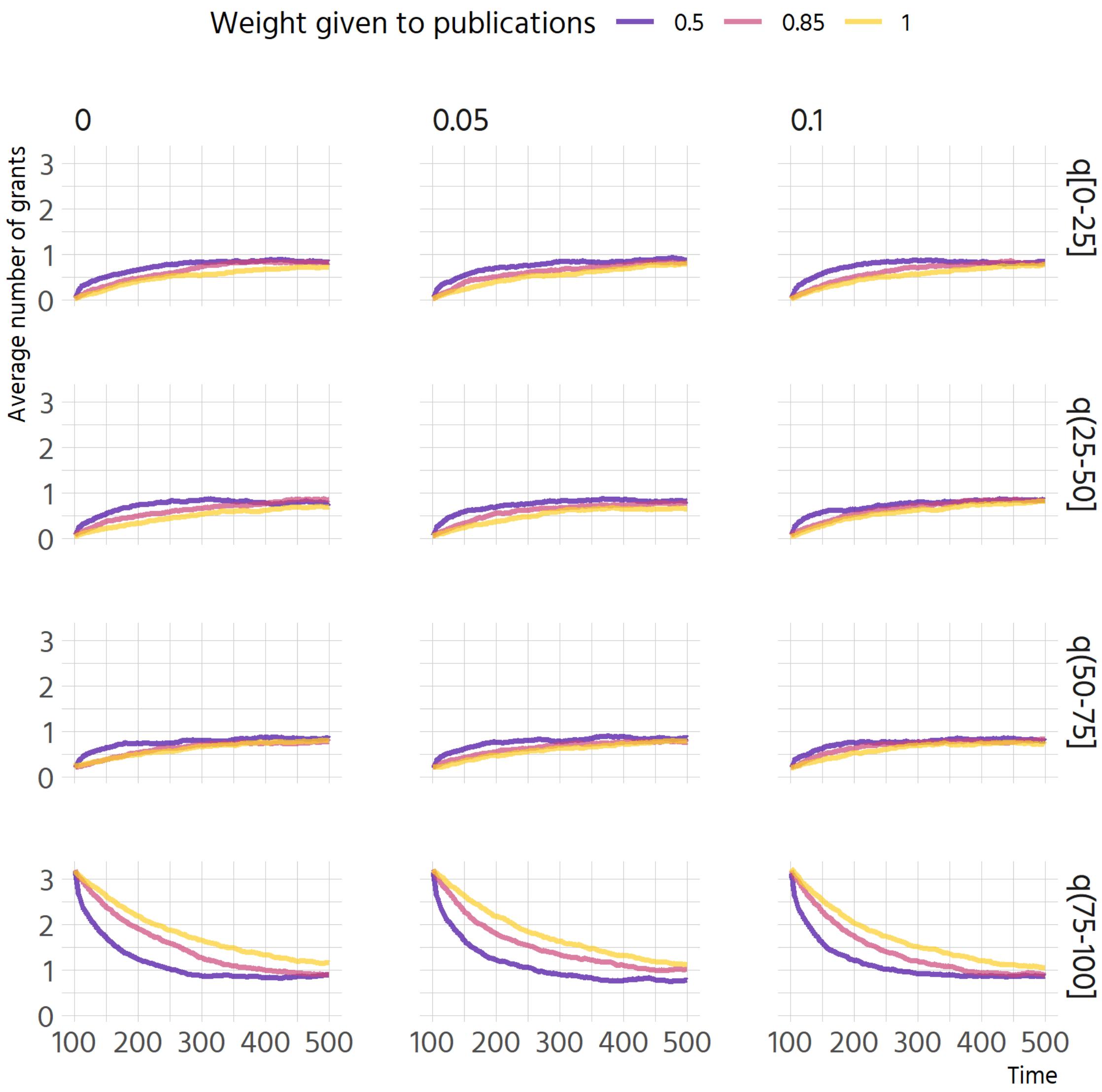

Figure 13: Average of the mean number of grants per group with individual learning. All long-term oriented agents. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the incentive setting of the funder. Aggregated from 100 runs.

4.3. Data sharing - Social learning

In contrast to the individual learning mechanism, agents change their behaviour considerably in adapting their strategies with social learning in our model. Below we describe results along the same dimensions as above: the diffusion of sharing behaviour, numbers of total datasets shared and publications produced, general equity of the system, and the impact on individual groups.

4.3.1.1. Diffusion of sharing behaviour

Equipped with the rationale for social learning described in section 3.5.2, agents are able to adapt their strategy to given structures of incentives and sharing costs. Figure 14 displays results from individual runs for long-term oriented agents. Results from myopic agents are presented in the appendix (Figure A1).

As in the previous section, Figure 14 contrasts sharing incentives (rows) with sharing costs (columns). Under no incentives and no costs (left bottom), there is no dominant strategy. Whether to share or not to share does not influence the publication success of groups. For other scenarios, adoption is slower or faster depending on the size of costs and incentives. For a scenario with clear incentives and no costs of sharing (top left; the history of data sharing receiving a weight of 50% in the funder's decision to fund (weight given to publications vs. datasets = 0.5)), adoption of data sharing is very quick. Vice versa, without incentives but with substantial costs (bottom right), adoption of non-sharing happens across most runs equally quickly.

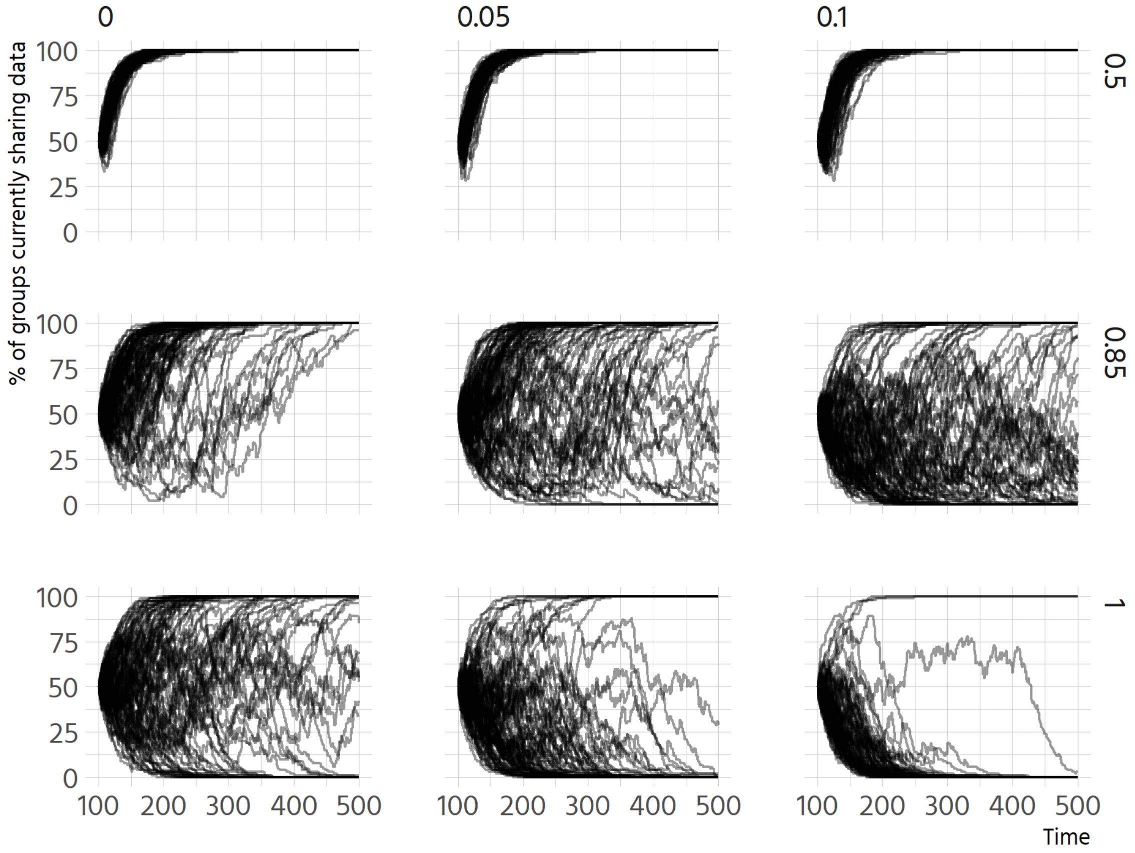

Figure 14: Share of research groups opting to share data - trajectories of long-term oriented agents. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

Contrasting agents with different time-orientations (Figure 15) reveals many interesting and complex patterns. While a system comprising solely myopic agents that compare themselves to their peers every round only weakly tends towards data sharing when there is no cost but some incentives (left middle), almost all long-term oriented agents or a mix of agents choose to share data quickly. The same holds true for scenarios with low costs (second column from left; cost = 0.05). Here, the longer time horizon of agents enables them to better observe successful behaviour among their peers, compared to all myopic and even mixed sets of agents.

In substantive terms, given a scenario of high costs to sharing data, considering prior data sharing behaviour to a moderate degree (15% of the funder's decision) is not able to offset the costs involved with sharing.

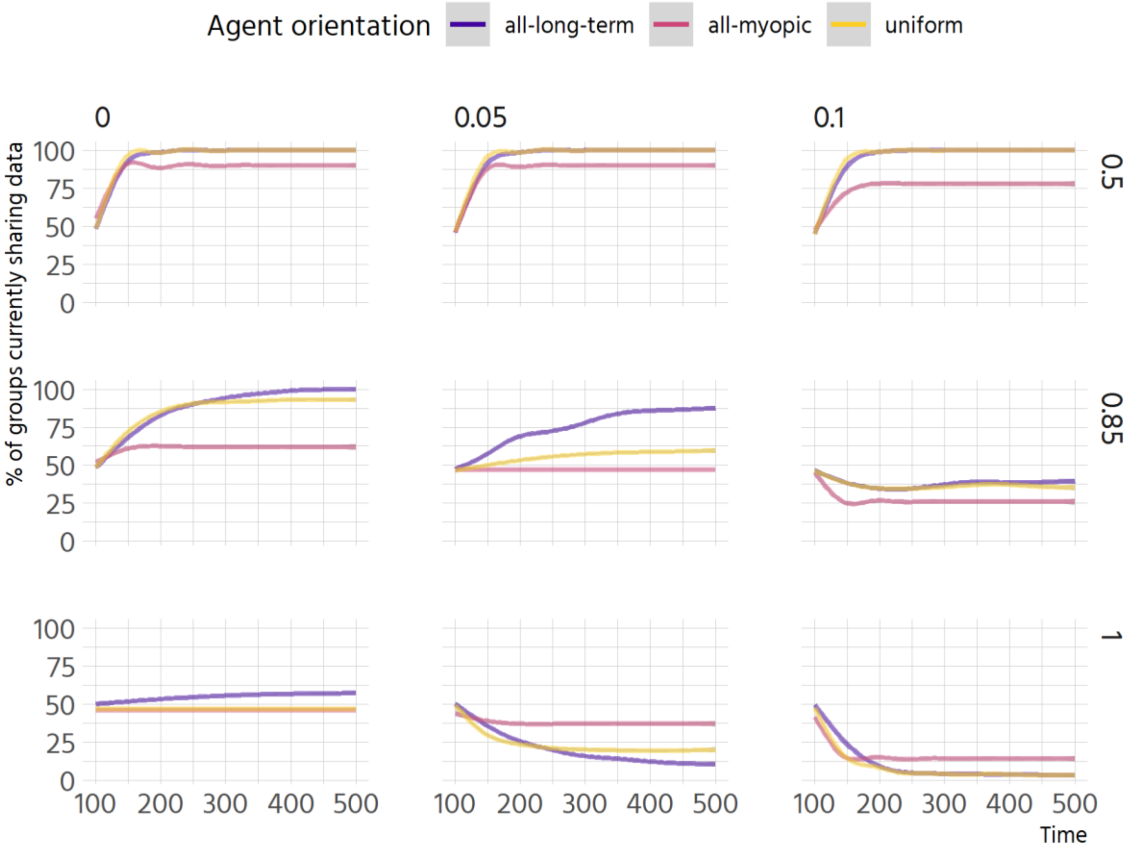

Figure 15: Share of research groups opting to share data - all time-orientations. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.3.2. Extent of data sharing

The extent of data being shared is, with social learning, closely linked to agents' decisions whether to share or not to share data (Figure 16). Overall, the numbers of available datasets are higher for scenarios with larger incentives, and lower for scenarios involving higher costs. In line with the above findings, moderate incentives and costs lead to higher numbers of data shared among long-term oriented agents, compared to solely myopic agents.

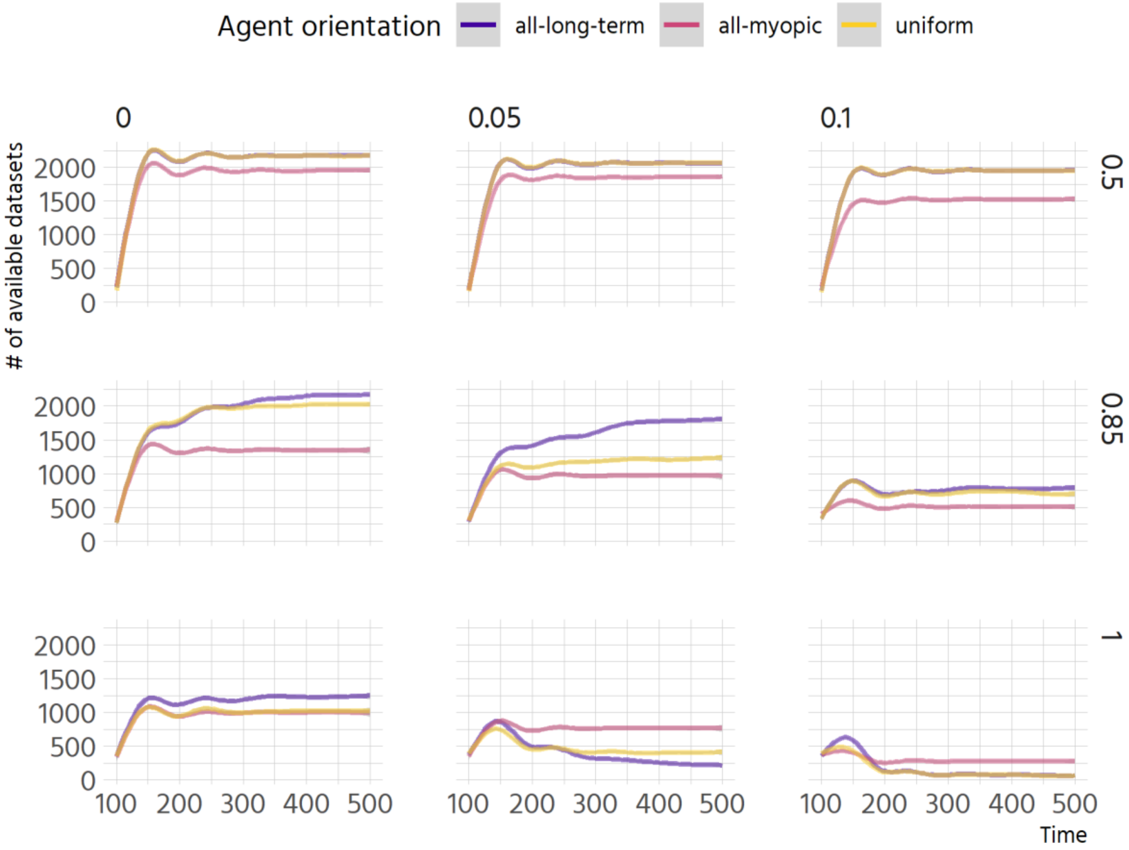

Figure 16: Number of available datasets - all time-orientations. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.3.3. Extent of publication activity

Considering the level of publication activity, we observe patterns that partly mirror the results from investigating distributions of produced datasets (Figure 17). Firstly, the level of current publications is constant and high for conditions that do not involve a cost associated with sharing data. Secondly, as agents either adapt to share or not to share data, the distribution of current publications also changes. This is driven by the fact that publications and datasets are related: more data sharing leads to lower publications overall, due to the cost inherent in making data available. With high costs (column-value = 0.1) and no incentives (row-value = 1), long-term oriented agents stop sharing quicker than others, and therefore produce more publications per round. Vice versa, with high costs (column-value = 0.1) and some incentives (row-value 0.85), long-term oriented agents tend to share data, thus producing fewer publications.

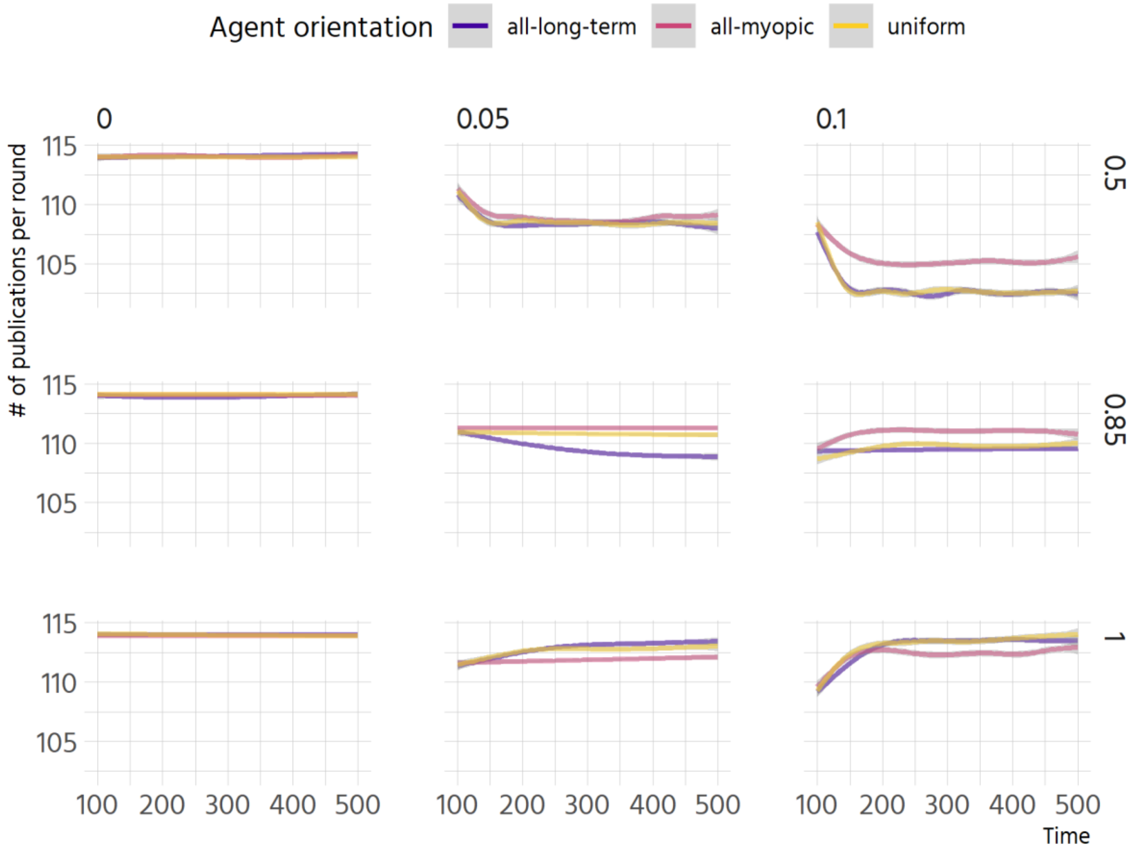

Figure 17: Publication productivity of groups. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.3.4. Equity of the system

When observing outcomes in terms of the equity of the distributions of datasets, grants, and publications, we find similar results for publications and grants, but substantially different results for the number of datasets produced per round (Figure 18). Considering grants and publications first, both exhibit stable distributions in terms of their inequality, almost irrespective of cost or incentive settings. Under high incentives for data sharing, the distribution of datasets produced per round quickly reaches a stable state of similar concentration like publications. In this scenario, all simulation runs end with everyone sharing data, for which reason the number of datasets equals the number of publications produced per round.

For scenarios with moderate to no incentives, we observe substantial differences in the level of concentration of datasets produced between different cost scenarios. With moderate incentives but no costs, concentration in the distribution of datasets per round declines, and reaches the equilibrium state at about round 300. However, in settings with higher costs, concentration declines more slowly, and in the case of assumed RDM-costs of 10%, initially rise and subsequently stay at higher levels. A tentative explanation for these effects is that under moderate incentives and comparatively high costs of sharing data, there are multiple successful strategies: some groups opt to share data, while others don't, and both strategies proliferate among a certain share of groups. Figure 15 indicates that this share is at about 50% of groups, with 10% costs and a weight of 15% towards data sharing success.

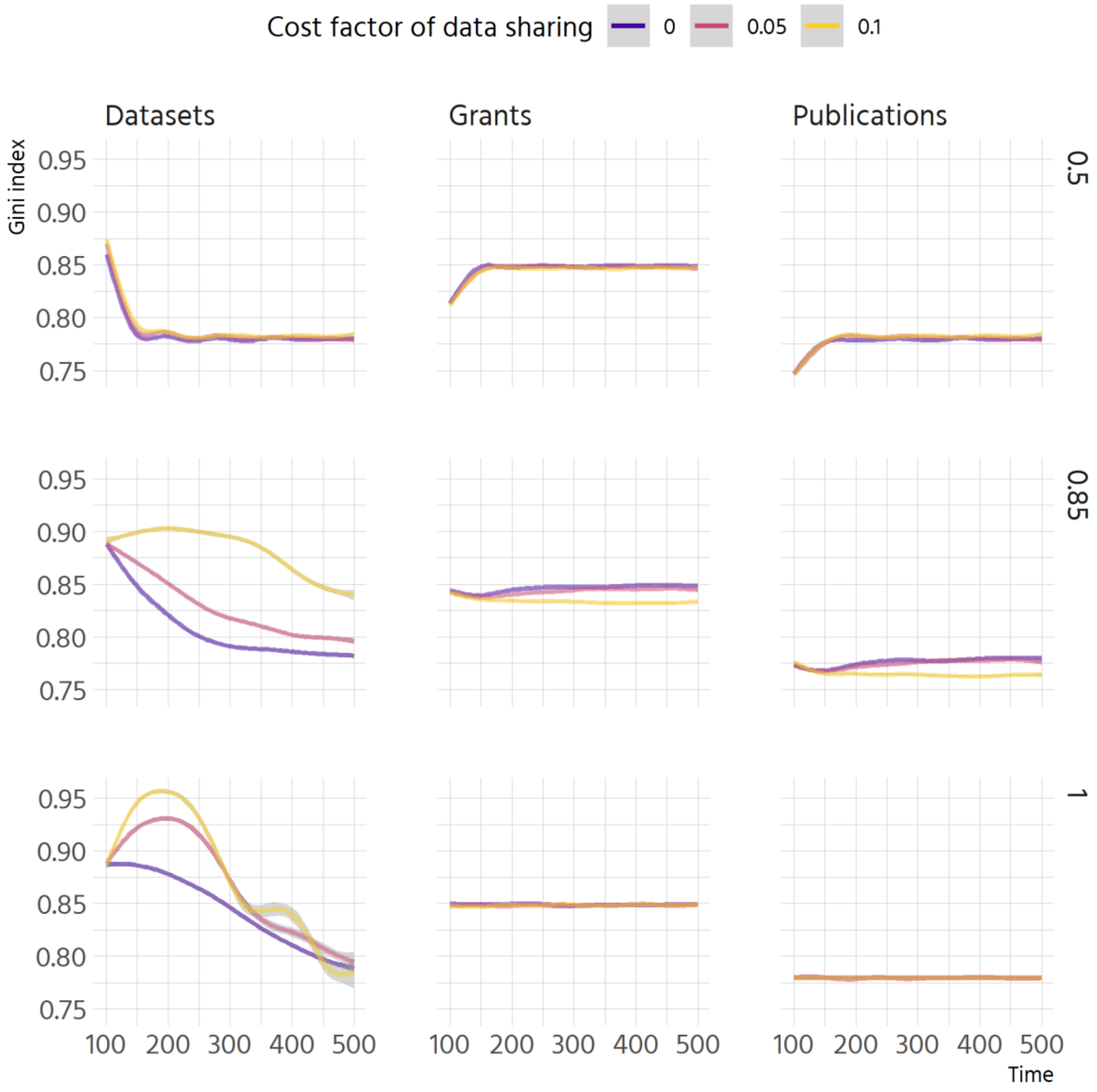

Figure 18: Gini indices of the distributions of datasets, grants and publications for decision rules based on social learning. All agents with long term orientation. The rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

4.3.5. Behaviour of groups by publication success

Finally, we analyse the choices of research groups on data sharing (using the same procedure to determine groups as described in section 4.2.5). The results of agents equipped with social learning are very different from those with individual learning. Under the current mechanism of social learning, groups adapt their behaviour at similar rates, regardless of their initial position in the distribution of publications (Figure 19). This result follows the logic of the learning mechanism, where groups choose 5 other groups *at random* and compare their behaviour. Under alternative scenarios, such as groups choosing peers to compare themselves to from network structures or based on the overall distribution of publications (e.g., choosing groups of similar success), different outcomes seem likely.

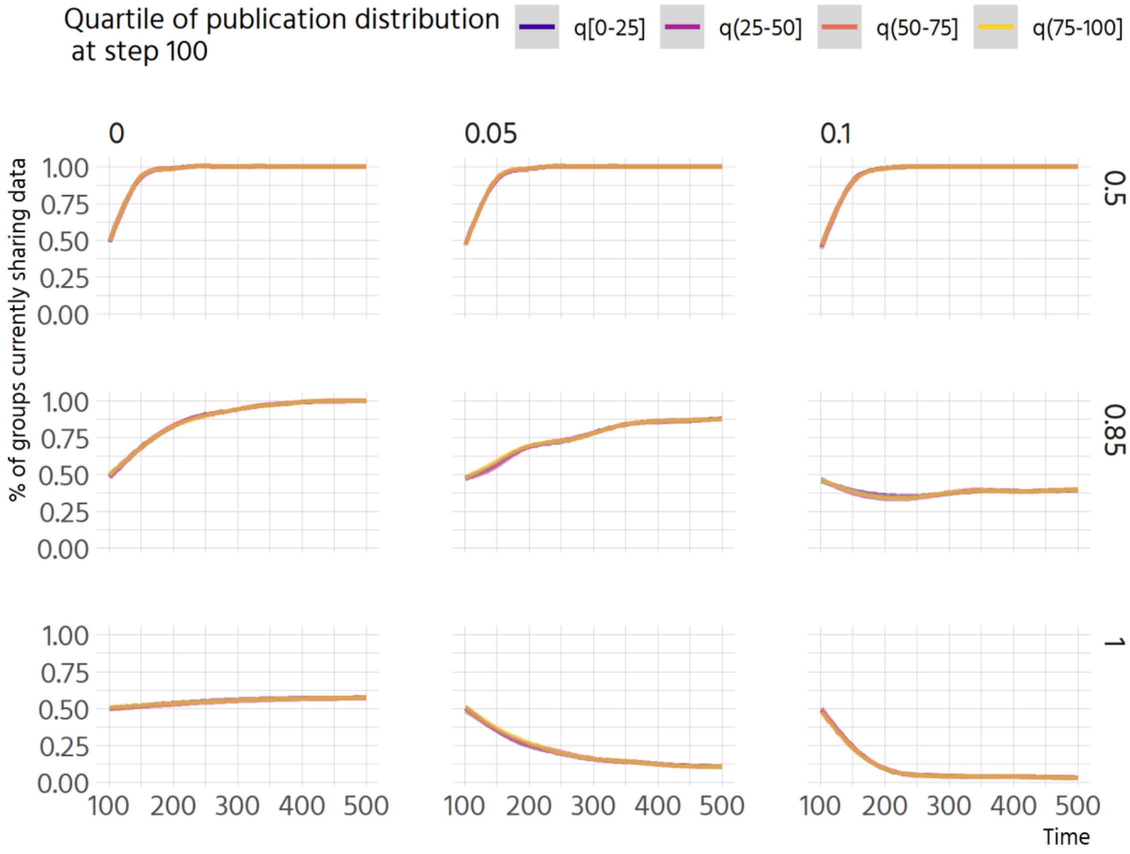

Figure 19: Proportion of groups sharing data by publication quartiles with social learning. All long-term oriented agents. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated from 100 runs.

Considering the average number of datasets produced per group, we find similar results to the model with individual learning (Figure 20). Higher cost of data sharing (columns) decreases the gap between the top quartile and all other groups, an effect that is substantially more pronounced in the model with social learning. The effect is strongest in settings where the system leads to uniform non-sharing. In settings where sharing persists, initially successful groups continue to produce substantially more datasets than all other groups, a gap that widens over time.

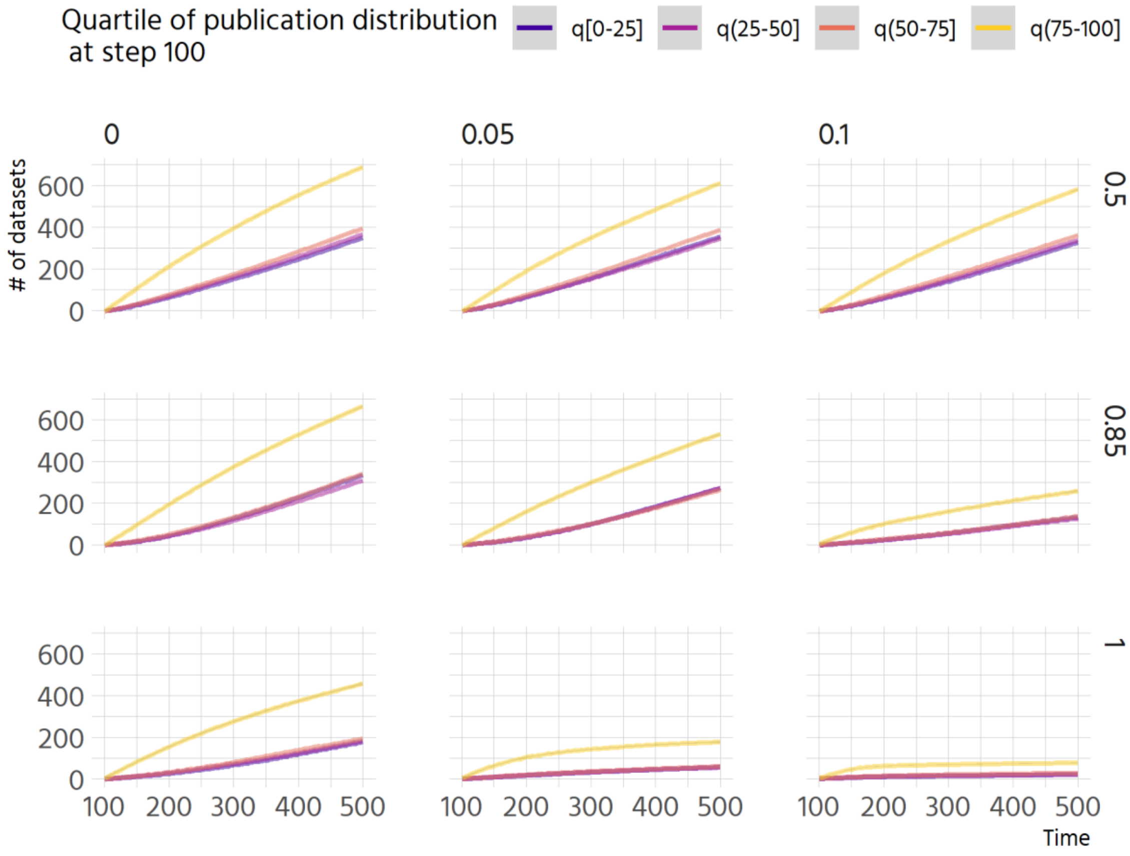

Figure 20: Mean number of total datasets per group with social learning. All long-term oriented agents. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated from 100 runs.

Considering the number of grants, we find similar but smaller differences than in the model with individual learning (Figure 21). Again, the average number of grants in the top group declines over time, and the rate of this decline varies with the weight given to datasets in deciding on funding. However, compared to the mechanism of individual learning, successful groups tend to retain their success longer, with the incentive regime having a smaller effect on the differential success of research groups.

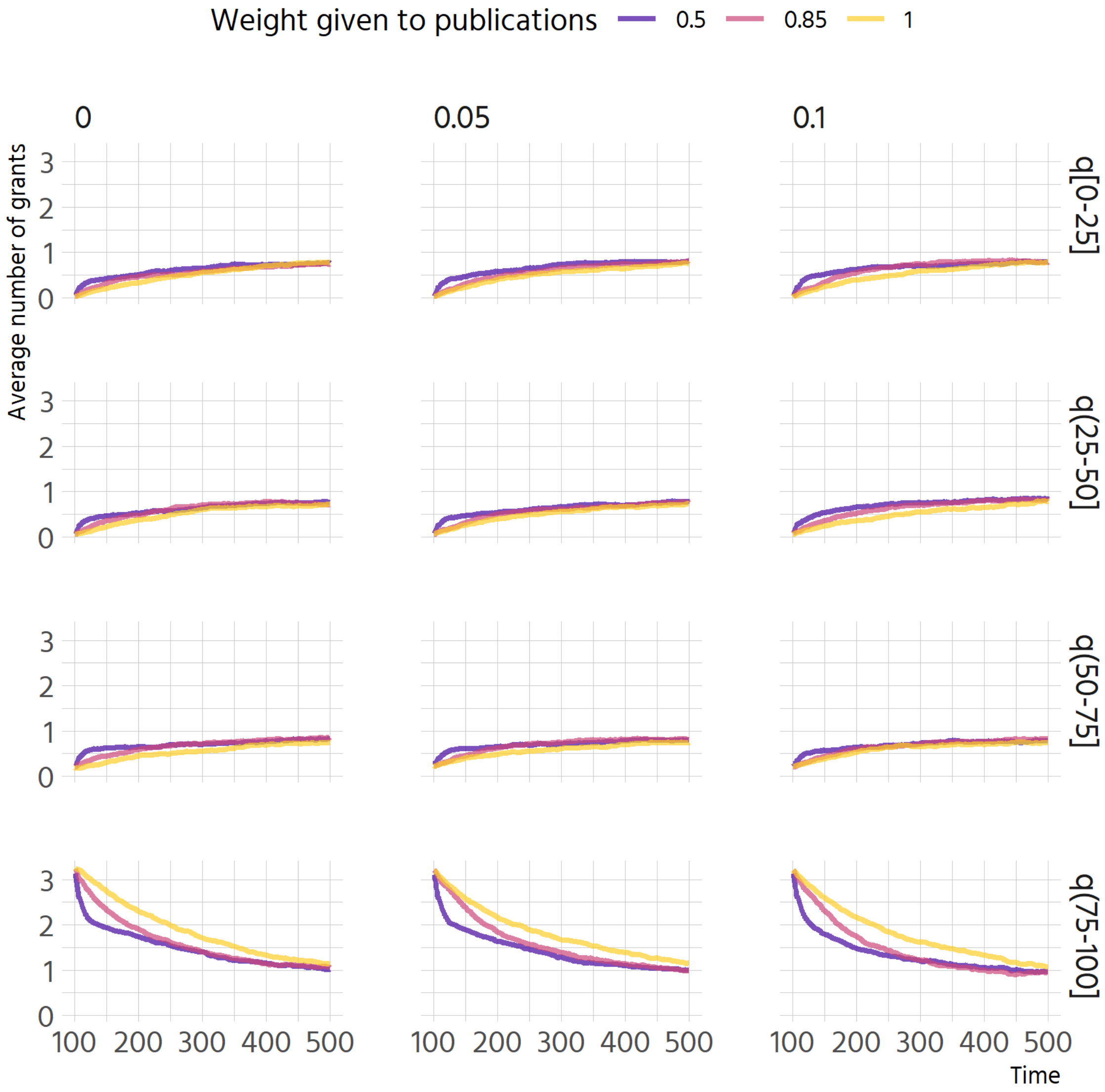

Figure 21: Average of the mean number of grants per group with social learning. All long-term oriented agents. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated from 100 runs.

5. Discussion

Data sharing has become an integral part of the Open Science agenda. It is seen as an important step towards more transparency and reproducibility, as well as bearing the potential to make scientific endeavours more effective through the reuse of data made available. Our report investigates important dynamics in the uptake of data sharing: how costs and incentives shape data sharing decisions amongst individual research groups, and how these individual decisions lead to aggregate outcomes at the system level. Furthermore, we analyse how these different scenarios relate to existing dynamics of cumulative advantage.

Various interventions to increase data sharing have previously been proposed, from recognizing data sharing via authorship, awarding badges to papers sharing data, to tracking data sharing via specific metrics (Woods & Pinfield, 2021). Based on their assessment of data sharing of articles across *Nature* and *Science*, Tedersoo et al. (2021) recommend creating real benefits for data sharing by taking it into account in applications for grants and academic positions. Any such intervention should be backed up by a solid understanding of expected outcomes and potential unintended consequences. As we have argued in section 2.3, ABM offers a coherent framework under which such interventions and their implications can be studied. With the DASH model we take a first step in improving our understanding of how interventions at the funder level might succeed to increase data sharing among academic research groups.

5.1. Time horizons of research groups - opportunistic versus strategic strategies

In our analysis, we compare agents with different time perspectives. We first consider myopic agents which act in an opportunistic fashion: revising their decision whether or not to share data every round, based on what other groups are doing or how well they themselves fare. We contrast these groups with agents that have a longer perspective. These differences in time perspective reflect two extremes of a continuum of strategies which are expected to be employed by academics. New and growing research groups might act opportunistically, to maximise their publication output, while more established groups might be more eager to take strategic decisions with longer time perspectives, being overall potentially more conservative.

Our results show that data sharing is taken up by groups with longer time horizons much more readily than their myopic counterparts, in settings where data sharing is beneficial (low to no costs but incentives for data sharing). In fact, under the model of individual learning (groups comparing their own success to their previous success), myopic groups are unable to adapt their strategies regarding whether or not to share data. The reason for this inability to adapt lies in the high contingency inherent with publishing and grant funding mechanisms. While some of our modelling conditions include substantial benefits to data sharing, these benefits do not become visible to research groups revising their decisions very frequently, since the chain of events from producing publications to sharing data to receiving grants which lead to further publications takes multiple rounds to take any effect. If funders were to incorporate incentives for sharing data, the effect on the overall system would thus strongly depend on the time orientation of groups in a given field of research. Assuming boundedly rational agents, Kwon and Motohashi (2021) argue that it is more likely researchers would make myopic decisions when considering data sharing, since they would be more aware of immediate positive effects such as gained credit, but limited in anticipating any detrimental negative effects which emerge later on.

5.2. Social learning

Among research groups following the rule of social learning, the whole system usually tends either towards sharing or towards not sharing data, i.e., a "consensus" emerges. Costs and incentives shape which settings predominantly lead to sharing and which prohibit sharing of data. Assuming moderate costs for RDM (5%) and moderate incentives (weight towards data sharing: 15%), sharing of data does not increase above 50% of all groups with solely myopic research groups. With long-term oriented agents, sharing happens in the majority of cases (about 90%), while with a uniform mix of agents, sharing remains close to 50%. Removing costs associated with RDM has a particularly strong effect on situations with a uniform mix of time-orientations, where sharing increases substantially to above 90%. Higher weights toward the history of data sharing lead to a quick uptake of data sharing across plausible cost conditions.

The approach of social learning is sensitive to the initial share of agents sharing data. Starting conditions with an equal percentage of agents sharing and not sharing data (as in our setup) allow us to examine the general tendencies of agents. More research should be conducted to incorporate more specific scenarios of relatively low initial rates of data sharing, which relate more closely to the current state of data sharing.

Our analysis therefore suggests two major conclusions: first, assuming rational agents and moderate costs of RDM (5%), substantial incentives are necessary to bring about data sharing among research groups. The magnitude of incentives has a clear effect on the speed of the uptake of data sharing. Second, reducing costs for data sharing increases the rate of uptake particularly for fields with heterogeneous time-orientations. Offsetting costs for data sharing therefore could therefore be an effective interim strategy for funders aiming at increasing rates of data sharing.

5.3. Practical considerations on how to fund data sharing

Proposals to incentivise the sharing of research data often also suggest that funders should cover the costs of appropriate research data management (Tedersoo et al., 2021). However, the implications of such an approach require careful attention. A first consideration is that the total budget available for funding research might not increase, so making RDM mandatory and the inherent costs fully refundable will likely reduce publication output of single projects. In a setting with no reuse of data, this would therefore lead to less publications, but higher rates of data sharing. This in turn could potentially lead to higher (computational) reproducibility (Peng & Hicks, 2021). However, as recent investigations into data sharing have shown (Hardwicke et al., 2018; Tedersoo et al., 2021), data might not in all cases be *actually* available for reuse, despite such claims in manuscripts. Scientists trying to maximise their returns would simply either state that data is available or upload data without detailed documentation. This would enable them to benefit from incentives while at the same time minimising resources spent on sharing data. Including simple metrics tracking data sharing in funding decisions would therefore be unlikely to provide sufficient incentives to drive the sharing of high-quality datasets which are a precondition for increased reproducibility and reuse.

5.4. Limitations and outlook

The DASH model presented in this deliverable rests on multiple assumptions, and subsequently has its limitations. One limitation pertains to the decision mechanism of the research groups, which is solely based on rationality. However, decisions whether or not to share data are likely more complex. First, it is reasonable to expect that some groups opt to share data irrespective of costs and incentives, simply because they believe it is the right thing to do (Fecher et al., 2015; Tenopir et al. 2015). This dimension could be incorporated into the decision mechanism, e.g. by building on Weber's distinction between *Wertrationalität* and

Zweckrationalität. Another approach would be to build on theories on the dissemination of culture, assuming that data sharing is more likely in disciplines which have close ties to other open approaches, such as open-source software. Further key forces that impinge on individual groups' decisions whether or not to share data are research funders, which increasingly mandate sufficient provisions to ensure data is shared as openly as possible (e.g., in Horizon Europe). Last but not least, there are fields in which the sharing of data is a precondition to publishing, like in astronomy. Some of these factors could be incorporated to map existing processes more clearly.

Our current model assumes one funder for reasons of simplicity, but it would be possible to add further funders to the system. This would allow us to probe how a mixture of incentive regimes would influence the speed of (non-)uptake of data sharing. Equally, the web of incentive regimes could be expanded by considering journals and their policies. Hardwicke et al. (2018) have shown that mandates by publishers to share data increase data sharing considerably. Extending the model in this regard would allow us to probe how the interaction between funder mandates/incentives and journal policies shape the uptake of data sharing, and the subsequent effect on equity in academia.

Restricting the scope of the model for the purpose of this deliverable involved a further simplifying assumption, which was to ignore the topic of data reuse. However, a key argument of stakeholders recommending researchers should share data is the purported potential this has for the subsequent reuse of data (Directorate-General for Research and Innovation, 2016). If one assumes that the costs for further publications are lower when relying on previously gathered data, it is reasonable to expect a higher efficiency across scientific disciplines. This line of reasoning at least partly relies on the notion that shared data are universally reusable, across contexts and time, which certainly is not true for all scientific disciplines (Reichmann, 2022). Nevertheless, investigating data sharing without considering reuse is only part of the picture. A revised version of the DASH model therefore will also consider data reuse and its wider implications for policies aimed at increasing rates of data sharing.

Another assumption concerns how research groups react to cost pressures inherent with data sharing. Our model assumes that the chances for publishing decrease as the cost for data sharing rises, since more time and effort is needed to prepare datasets, which in turn is not available to work on publications. This rests on the simplifying assumptions that all publications are of equal quality. However, it would also be reasonable to assume that groups react to cost pressures by writing shorter publications which cover fewer experiments. It is unclear whether more but shorter publications, or fewer but longer publications would be more efficient for science as a whole. Either way, journal publications have become substantially shorter since the 1950s, especially in the period 1950-1980, and since the year 2000 (Fire & Guestrin, 2019).

Finally, a key area for extending the DASH model would be to incorporate network structures. In the current implementation, interaction between the agents is either not modelled (individual decision rule) or based on chance (social decision rule). However, research groups are embedded in a network of collaboration and competition. It is therefore reasonable to assume that practices such as data sharing diffuse along lines of interaction between groups, based on collaboration or simply on being adjacent in the specific topics which groups investigate. Incorporating such dynamics into the model would enable us to analyse how more central actors influence the overall dynamic of the system, and which effect could be expected from targeting such central hubs in the network of academic endeavours.

The DASH model is a first attempt at modelling policy interventions regarding data sharing. The evidence presented in this deliverable shows that ABM offers a coherent methodology to tackle pressing questions of how varying incentives for data sharing shape its uptake and lead to potential repercussions for equity within academia. Our upcoming work in D6.4 on guidelines and recommendations for equitable implementations of OS will put forward a range of evidence-based recommendations. Assessing proposed interventions targeting the implementation of OS should always be conducted using suitable methods to investigate their efficacy and potential side-effects. ABM offers a valuable toolbox in this regard, and should be employed for such tasks in the future.

6. Conclusions

Open Science aims at reshaping crucial areas of research: how research is being conducted, how it is communicated, whether and how research artefacts are shared, how scientists are being rewarded and promoted, and how research engages with industry and the wider society. Part of this change are a host of stakeholders: research funders, research performing organisations, governments, companies (be it publishers or firms relying on scientific knowledge), researchers themselves, and the wider society. All these actors influence how research is being conducted, with a particularly direct path being policies put forward by research funders, publishers, and research performing organisations. Any such interventions should be backed up by a solid understanding of expected outcomes and potential unintended consequences. As we have argued throughout this deliverable, ABM offers a coherent framework under which such interventions and their implications can be studied.

Considering a key area of OS, the sharing of research data, our DASH model takes a first step in improving our understanding of how interventions at the funder level might succeed to increase data sharing among academic research groups. We first investigated the uptake of data sharing assuming rational agents who compare their current with their past success. Our analysis suggests that reasonable incentives for data sharing, such as considering an applications history data sharing to an equal amount in funding decisions as their publication history, are not sufficient to incur wide-spread adoption of data sharing. The contingencies inherent to the (stylised) process of research (funding → research → publications → funding) prohibit agents from adjusting their behaviour.

Considering agents that learn by imitating their peers, our investigations provide two key findings: first, assuming rational agents and moderate costs of RDM (5%), substantial incentives are necessary to bring about data sharing among research groups. The magnitude of incentives has a clear effect on the speed of the uptake of data sharing. Second, reducing costs for data sharing increases the rate of uptake particularly for fields where agents exhibit heterogeneous time-orientations (some of them revising their decisions very frequently, others revising less frequently).

Our model and analysis provide a first step in analysing proposed paths of action to encourage the sharing of research data. There are manifold options for extending the model, improving its accuracy, expanding its scope, and widening the types of policies considered. Nevertheless, the DASH model provides a solid illustration of how ABM can be leveraged to inform policy-making processes and strengthens the evidence upon which the upcoming ON-MERRIT policy recommendations (Deliverable 6.4) will be based.

7. References

Abbott, A. D. (2004). *Methods of discovery: Heuristics for the social sciences*. W.W. Norton & Co.

Ahrweiler, P. (2017). Agent-based simulation for science, technology, and innovation policy. *Scientometrics*, *110*(1), 391–415. https://doi.org/10.1007/s11192-016-2105-0

Ahrweiler, P., Schilperoord, M., Pyka, A., & Gilbert, N. (2015). Modelling research policy: Ex-ante evaluation of complex policy instruments. *JASSS*, *18*(4). Scopus. https://doi.org/10.18564/jasss.2927

Alexander, J. C., Giesen, B., Münch, R., & Smelser, N. J. (Eds.). (1987). *The Micro-Macro Link* (2. print). Univ. of California Press.

Bernius, S., Hanauske, M., Dugall, B., & König, W. (2013). Exploring the effects of a transition to open access: Insights from a simulation study. *Journal of the American Society for Information Science and Technology*, *64*(4), 701–726. https://doi.org/10.1002/asi.22772

Bollen, J., Crandall, D., Junk, D., Ding, Y., & Börner, K. (2017). An efficient system to fund science: From proposal review to peer-to-peer distributions. *Scientometrics*, *110*(1), 521–528. https://doi.org/10.1007/s11192-016-2110-3

Borgman, C. L., & Bourne, P. E. (2021). Why it takes a village to manage and share data. *ArXiv:2109.01694 [Cs, Econ, q-Fin]*. http://arxiv.org/abs/2109.01694

Bruch, E., & Atwell, J. (2015). Agent-Based Models in Empirical Social Research. *Sociological Methods & Research*, *44*(2), 186–221. https://doi.org/10.1177/0049124113506405

Brugnach, M., Tagg, A., Keil, F., & de Lange, W. J. (2007). Uncertainty matters: Computer models at the science-policy interface. *Water Resources Management*, *21*(7), 1075–1090. Scopus. https://doi.org/10.1007/s11269-006-9099-y

Bullock, S. (2016). Alife as a Model Discipline for Policy-Relevant Simulation Modelling: Might “Worse” Simulations Fuel a Better Science-Policy Interface? (Extended Abstract). *Artificial Life XV: Proceedings of The Fifteenth International Conference on the Synthesis and Simulation of Living Systems*, 28–29. https://doi.org/10.1162/ecal_a_0010

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Cole, N. L., Reichmann, Stefan, Tony, R.-H., & Bernhard, W. (2021). *ON-MERRIT D5.3 Networks of Engagement in Deliberative Policy-making: Expert Reflections on Barriers to Participation*. https://doi.org/10.5281/ZENODO.5550533

de Marchi, S., & Page, S. E. (2014). Agent-Based Models. *Annual Review of Political Science*, *17*(1), 1–20. https://doi.org/10.1146/annurev-polisci-080812-191558

Directorate-General for Research and Innovation (Ed.). (2016). *Open innovation, open science, open to the world: A vision for Europe*. Publications Office of the European Union.

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *PLOS ONE*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience*, *8*(6). https://doi.org/10.1093/gigascience/giz053

Franceschini, F., & Maisano, D. (2011). Structured evaluation of the scientific output of academic research groups by recent h-based indicators. *Journal of Informetrics*, *5*(1), 64–74. https://doi.org/10.1016/j.joi.2010.08.003

Gilbert, N. (2008). *Agent-Based Models*. SAGE Publications, Inc. https://doi.org/10.4135/9781412983259

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Higginson, A. D., & Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, *14*(11), e2000995. https://doi.org/10.1371/journal.pbio.2000995

Johnson, J. A. (2018). Open data, big data, and just data. *Public Administration and Information Technology*, *33*, 23–49. Scopus. https://doi.org/10.1007/978-3-319-70894-2_2

Kwon, S., & Motohashi, K. (2021). Incentive or disincentive for research data disclosure? A large-scale empirical analysis and implications for open science policy. *International Journal of Information Management*, *60*, 102371. https://doi.org/10.1016/j.ijinfomgt.2021.102371

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts: with a new postscript and index by the authors* (1. Princeton paperback printing). Princeton University Press.

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

Lempert, R. (2002). Agent-based modeling as organizational and public policy simulators. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7195–7196. https://doi.org/10.1073/pnas.072079399

Lorig, F., Johansson, E., & Davidsson, P. (2021). Agent-Based Social Simulation of the Covid-19 Pandemic: A Systematic Review. *Journal of Artificial Societies and Social Simulation*, *24*(3), 5.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *ELife*, *5*, e16800. https://doi.org/10.7554/eLife.16800

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, *578*(7796), 491–491. https://doi.org/10.1038/d41586-020-00505-7

Morgan, M. S., & Morrison, M. (1999). Models as mediating instruments. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators: Perspectives on natural and social sciences*. Cambridge University Press.

Murray, E. J., Robins, J. M., Seage, G. R., Freedberg, K. A., & Hernán, M. A. (2017). A Comparison of Agent-Based Models and the Parametric G-Formula for Causal Inference. *American Journal of Epidemiology*, *186*(2), 131–142. https://doi.org/10.1093/aje/kwx091

Peng, R. D., & Hicks, S. C. (2021). Reproducible Research: A Retrospective. *Annual Review of Public Health*, *42*, 79–93. https://doi.org/10.1146/annurev-publhealth-012420-105110

Price, D. J. D. S. (1963). Little Science, Big Science. In *Little Science, Big Science*. Columbia University Press. https://doi.org/10.7312/pric91844

Pronk, T. E., Wiersma, P. H., Weerden, A. v., & Schieving, F. (2015). A game theoretic analysis of research data sharing. *PeerJ*, *2015*(9). Scopus. https://doi.org/10.7717/peerj.1242

Reichmann, S. (2022). Data sharing for global, unspecified reuse. *SocArXiv*. https://doi.org/10.31235/osf.io/tbex2

Ross-Hellauer, T., Reichmann, S., Cole, N. L., Fessl, A., Klebel, T., & Pontika, N. (2021). *Dynamics of Cumulative Advantage and Threats to Equity in Open Science—A Scoping Review*. SocArXiv. https://doi.org/10.31235/osf.io/d5fz7

Scheer, D. (2017). Between knowledge and action: Conceptualizing scientific simulation and policy-making. In *The Science and Art of Simulation I: Exploring—Understanding—Knowing* (pp. 103–118). Scopus. https://doi.org/10.1007/978-3-319-55762-5_8

Scheliga, K., & Friesike, S. (2014). Putting open science into practice: A social dilemma? *First Monday*. https://doi.org/10.5210/fm.v19i9.5381

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D., & Ioannidis, J. P. A. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biology*, *19*(3), e3001107. https://doi.org/10.1371/journal.pbio.3001107

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Squazzoni, F. (2012). *Agent-Based Computational Sociology* (1st edition). Wiley.

Squazzoni, F., Polhill, J. G., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., Chappin, É., Borit, M., Verhagen, H., Giardini, F., & Gilbert, N. (2020). Computational Models That Matter During a Global Pandemic Outbreak: A Call to Action. *Journal of Artificial Societies and Social Simulation*, *23*(2), 10.

Taghikhah, F., Filatova, T., & Voinov, A. (2021). Where Does Theory Have It Right? A Comparison of Theory-Driven and Empirical Agent Based Models. *Journal of Artificial Societies and Social Simulation*, *24*(2), 4. https://doi.org/10.18564/jasss.4573

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. https://doi.org/10.1038/s41597-021-00981-0

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Van Eck, N. J. (2021). *CWTS Leiden Ranking 2020* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4745545

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ*, *5*, e3208. https://doi.org/10.7717/peerj.3208

Watts, C., & Gilbert, N. (2011). Does cumulative advantage affect collective learning in science? An agent-based simulation. *Scientometrics*, *89*(1), 437–463. Scopus. https://doi.org/10.1007/s11192-011-0432-8

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Woods, H. B., & Pinfield, S. (2021). *Incentivising research data sharing: A scoping review* (6:355). Wellcome

Open Research. https://doi.org/10.12688/wellcomeopenres.17286.1

8. Annex

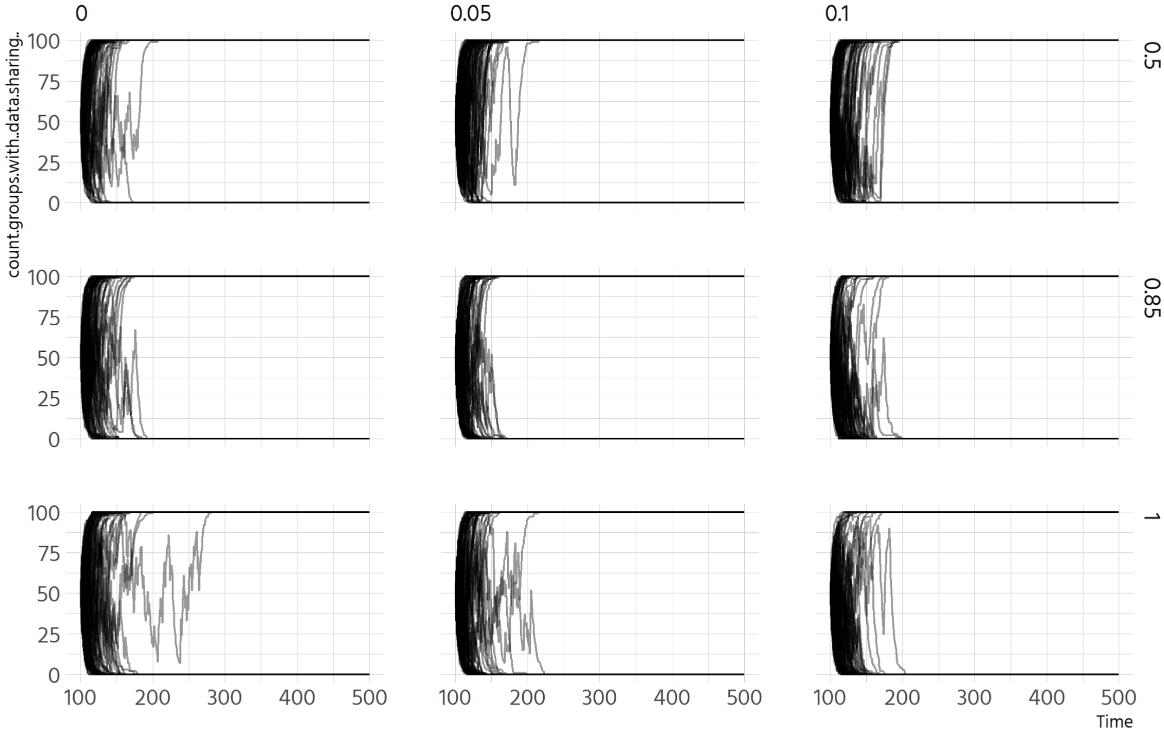

Figure A1: Share of research groups opting to share data - trajectories of myopic agents. The columns (0, 0.05, 0.1) refer to the cost factor inherent with sharing data, while the rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.

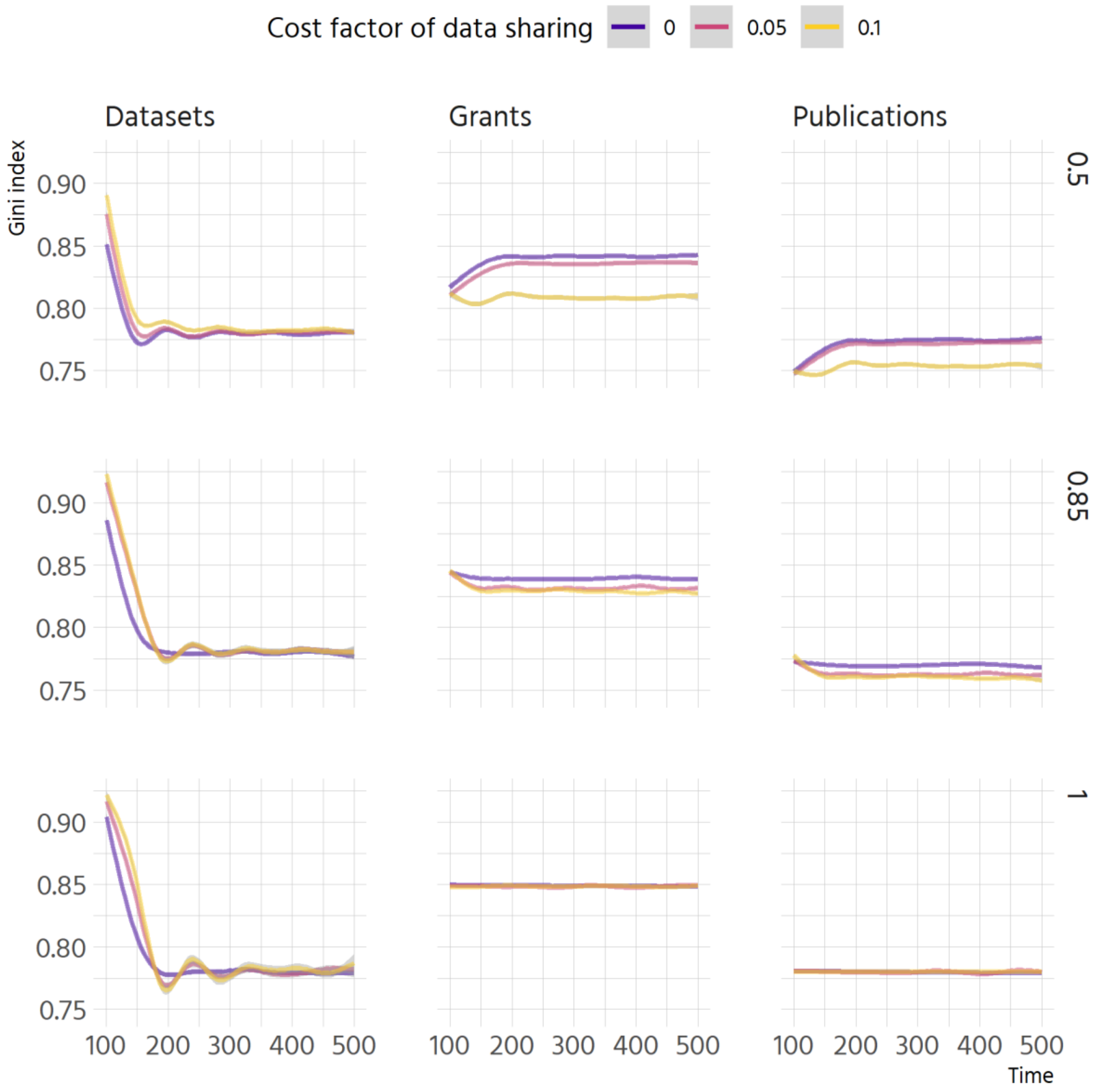

Figure A2: Gini indices of the distributions of datasets, grants and publications for decision rules based on social learning. All agents with myopic orientation. The rows (1, 0.85, 0.5) refer to the incentive setting of the funder. Aggregated results of 100 runs per condition.