The Large-Scale, Systematic and Iterated Comparison of Agent-Based Policy Models

Mike Bithell[1], Edmund Chattoe-Brown[2][0000-0001-8232-6896] and Bruce Edmonds[3][0000-0002-3903-2507]

[1] Department of Geography, Cambridge University
mike.bithell@geog.cam.ac.uk
[2] School of Media, Communication and Sociology, University of Leicester
ecb18@leicester.ac.uk
[3] Centre for Policy Modelling, Manchester Metropolitan University
bruce@edmonds.name

Abstract. Vital to the increased rigour (and hence reliability) of Agent-based modelling are various kinds of model comparison. The reproduction of simulations is an essential check that models are as they are described. Here we argue that we need to go further and carry out large-scale, systematic and persistent model comparison – where different models of the same phenomena are compared against standardised data sets and each other. Lessons for this programme can be gained from the Model Intercomparison Projects (MIP) in the Climate Community and elsewhere. The benefits, lessons and particular difficulties of implementing a similar project in social simulation are discussed, before sketching what such a project might look like. It is time we got our act together!

Keywords: model inter-comparison, policy modelling, replication, reproduction, IPCC, ABM, verification, COVID-19

1 Model Comparison, Alignment, Replication and Reproduction

The replicability of real-world experiments has long been a cornerstone of science. However, the recent "replication crisis" in psychology (for an account see [19.]) has brought this issue to the fore. In that context, replication means that if you followed the whole reported process of an experiment – selecting subjects, doing the experiment, analysing the results, etc. – you would come to the same conclusions as the original. A set of overlapping ideas has been imported into the world of Agent-Based Modelling (ABM) [2., 5., 8.], but the introduction of an artefact (the simulation model) adds a new stage to the sequence linking ideas to evidence (which implies some new distinctions). In computer science there are two sides to checking any code: (a) the *verification* of the code – that the code complies with its specification and (b) the *validation* of the code – that the code achieves its goals when implemented and used in practice. In ABM, the code is the computer simulation, so these translate as

two processes: (a) checking the code complies to its description including whether there are any hidden assumptions, bugs, etc. and (b) checking that the model relates to observed data in a way that supports the conclusions drawn from the modelling exercise. Since the replication crisis in psychology, there has been more focus on these kinds of processes, resulting in new terminology. These days we distinguish between *reproduction* which means re-coding the ABM from its description and checking one gets essentially the same results (to check (a)) and *replication* which also checks the validation (b). Thus, whilst [2.] talks about 'aligning' simulation models and [5.] talks about 'replicating' simulation models, under modern terminology these would both be 'reproducing' simulations. Axtell et al. [2.] is the first reported case of model reproduction we know of, showing that reproducing even simple models uncovers assumptions and differences. Edmonds and Hales [5.] independently reproduced a simple published model [24.] into two, very different, programming languages and then checked these two reproductions against each other (as well as the reported results) and showed that the original authors had a flawed understanding of their own model, changing its interpretation. Hales [15.] argues for a system of "replication-first" publication, whereby the model reported on is independently reproduced before it is published in order to ensure the completeness and reliability of reports on ABM research. Chattoe-Brown et al. [8.] argues that the reproduction of models is an essential check before any policies should be based on their results.

One important reason for these terminological challenges, which suggests that further analysis will be needed, is that models (and science generally) differ in the extent to which publication can serve as an effective summary of the actual research process. A publication, like a model, is to some extent an artefact to which the issues of verification and validation apply. Psychological experiments and "toy" ABM are such that they might be accurately reported in a standard article for the purposes of model duplication (though there is also a problem about whether this reporting is adequately performed in practice – see [25.]). If code is made available (which it often isn't) then a replication attempt can be directly checked. But for complicated models involving large scale data it may simply not be realistic to "make the model again" and other methods of establishing correspondence may be necessary [5].

Model alignment, replication and reproduction are important examples of a wider class of model-to-model analysis [16.]. This includes two important cases: *meta-modelling* and *model comparison*. Meta-modelling is where one models an existing model using another – for example, one might model a complex, descriptive ABM with a much simpler individual-based model in order to determine what is and is not essential for producing the same dynamics (e.g. [18.]). Model comparison occurs when we relate the outputs of intentionally different models – e.g. in their ability to fit certain target sets of data. In this paper we are talking about the latter category of model comparison.

2 The Need for Model Intercomparison

The recent COVID-19 crisis has led to a surge of new model development and a renewed interest in the use of models as policy tools. While this is in some senses welcome, the sudden appearance of many new models presents a problem in terms of their assessment, the appropriateness of their application and reconciliation of any differences in outcome. Even if they appear similar, their underlying assumptions may differ, their initial data might not be the same, policy options may be conceptualised in different ways, stochastic effects explored to varying extents, and model outputs presented in any number of different forms. Modelled processes may be disjoint, with some considering social processes, but others focussed purely on epidemiology, for example, and reported outputs may not even be of the same type, with no obvious way to compare them. As a result, it can be unclear what aspects of variations in output between models are the results of mechanistic, parameter or data differences. Any comparison between models is rendered difficult by differences in experimental design and selection of output measures.

If we wish to do better, we suggest that a more formal approach to making comparisons between models would be helpful. However, it appears that this is not commonly undertaken in most fields in a systematic and persistent way, except for the field of climate change, and closely related fields such as pollution transport or economic impact modelling (although efforts are underway to extend such systematic comparison to ecosystem models [28., 27.]). Examining the way in which this is done for climate models may therefore prove instructive.

3 Some Existing Model Comparison Projects

2.1 Model Intercomparison Projects (MIP) in the Climate Community

Formal intercomparison of atmospheric models goes back at least to 1989 [11.], with the first atmospheric model inter-comparison project (AMIP), initiated by the World Climate Research Programme. By 1999 this had contributions from all significant atmospheric modelling groups, providing standardised time-series of over 30 model variables for one particular historical decade of simulation, with a standard experimental setup. Comparisons of model mean values with available data helped to reveal overall model strengths and weaknesses: no single model was best at simulating all aspects of the atmosphere, with accuracy varying greatly between simulations. The model outputs also formed a reference base for further inter-comparison experiments including targets for model improvement and reduction of systematic errors, as well as a starting point for improved experimental design, software and data management standards and protocols for communication and model intercomparison. This led to AMIPII and, subsequently, to a series of Climate model inter-comparison projects (CMIP) beginning with CMIP I in 1996. The latest iteration (CMIP 6) is a collection of 23 separate model intercomparison experiments covering atmosphere, ocean, land surface, geo-engineering, and the paleoclimate. This collection is aimed at the upcom-

ing 2021 IPCC process (AR6). Participating projects go through an endorsement process for inclusion, (a process agreed with modelling groups), based on 10 criteria designed to ensure some degree of coherence between the various models – a further 18 MIPS are also listed as currently active [6.]. Groups contribute to a central set of common experiments covering the period 1850 to the near-present. An overview of the process can be found in [10.].

The current structure includes a set of three overarching questions covering the dynamics of the earth system, model systematic biases and understanding possible future change under uncertainty. Individual MIPS may build on this to address one or more of a set of 7 "grand science challenges" associated with the climate. Modelling groups agree to provide outputs in a standard form, obtained from a specified set of experiments under the same design, and to provide standardised documentation to go with their models. Originally (up to CMIP 5), outputs were then added to a central public repository for further analysis, however the output grew so large under CMIP6 that now the data is held dispersed over repositories maintained by separate groups.

2.2 Other Examples

Firstly, an informal network collating models across more than 50 research groups has already been generated as a result of the COVID-19 crisis – *the Covid Forecast Hub* [7.]. This is run by a small number of research groups collaborating with the US Centre for Disease Control and is strongly focussed on epidemiological aspects. Participants are encouraged to submit weekly forecasts, and these are integrated into a data repository and can be visualised on the website – viewers can look at forward projections, along with associated confidence intervals and model evaluation scores, including those for an ensemble of all models. The focus on forecasts in this case arises out of the strong policy drivers for the current crisis, but the main point is that it is possible to immediately view measures of model performance and to compare the different model types: one clear message that rapidly becomes apparent is that many of the forward projections have 95% (and at some times, even 50%) confidence intervals for incident deaths that more than span the full range of the past historic data. The benefit of comparing many different models in this case is apparent, as many of the historic single-model projections diverge strongly from the data (and the models most in error are not consistently the same ones over time), although the ensemble mean tends to be better.

As a second example, one could consider the Psychological Science Accelerator (PSA) [20., 23.]. This is a collaborative network set up with the aim of addressing the "replication crisis" in psychology: many previously published results in psychology have proved problematic to replicate as a result of small or non-representative sampling or use of experimental designs that do not generalise well or have not been used consistently either within or across studies. The PSA seeks to ensure accumulation of reliable and generalisable evidence in psychological science, based on principles of inclusion, decentralisation, openness, transparency and rigour. The existence of this network has, for example, enabled the reinvestigation of previous experiments but with much larger and less nationally biased samples (e.g. [17.]).

3 The Benefits of the Intercomparison Exercises and Collaborative Model Building

More specifically, long-term intercomparison projects help to achieve the following.

- *Build on past effort*. Rather than modellers re-inventing the wheel (or building a new framework) with each new model project, libraries of well-tested and documented models, with data archives, including code and experimental design, would allow researchers to more efficiently work on new problems, building on previous coding effort
- *Aid replication.* Focussed long term intercomparison projects centred on model results with consistent standardised data formats would allow new versions of code to be quickly tested against historical archives to check whether expected results could be recovered and where differences might arise, particularly if different modelling languages and approaches (compartmental, system dynamics, ABM) were being used
- *Help to formalise*. While informal code archives can help to illustrate the methods or theoretical foundations of a model, intercomparison projects help to understand which kinds of formal model might be good for particular applications, and which can be expected to produce helpful results for given desired output measures
- *Build credibility*. A continuously updated set of model implementations and assessment of their areas of competence and lack thereof (as compared with available datasets) would help to demonstrate the usefulness (or otherwise) of ABM as a way to represent social systems
- *Influence Policy* (where appropriate). Formal international policy organisations such as the IPCC or the more recently formed IPBES are effective partly through an underpinning of well tested and consistently updated models. As yet it is difficult to see whether such a body would be appropriate or effective for social systems, as we lack the background of demonstrable accumulated and well tested model results.

4 Lessons for ABM?

What might we be able to learn from the above, if we attempted to use a similar process to compare ABM policy models?

1. The projects started small and grew over time: it would not be necessary, for example, to cover all possible ABM applications at the outset. On the other hand, the latest CMIP iterations include a wide range of different types of model covering many different aspects of the earth system, so that the breadth of possible model types need not be seen as a barrier. There are several good arguments (current interest, policy relevance, intellectual challenge, plentiful “raw material”) for using pandemic policy as a demonstrator for this approach which could then – or in parallel – be expanded to other “significant” areas of ABM like opinion dynamics and land use modelling.

2. The climate inter-comparison project has persisted for about 30 years – over this time many models have come and gone, but the history of inter-comparisons allows for an overview of how well these models have performed over time – data from the original AMIP I models is still available on request, supporting assessments concerning long-term model improvement.
3. Although climate models are complex – implementing a variety of different mechanisms in different ways – they can still be compared by use of standardised outputs, and at least some (although not necessarily all) have been capable of direct comparison with empirical data. Thus the approach proposed here (unlike strict replication) is not limited to the analysis of models simple enough to be well described in a single article.
4. An agreed experimental design and public archive for documentation and output that is stable over time is needed; this needs to be done via a collective agreement among the modelling groups involved so as to ensure a long-term buy-in from the community as a whole, so that there is a consistent basis for long-term model development, building on past experience. This may mean a degree of compromise between groups on what models include and report, so as to avoid problems models that are in fact conceptually incompatible, or with outputs that are incommensurable.

The community has already established a standardised form of documentation in the ODD protocol and is working on further standardisation as methodology develops, in empirical modelling for example [25]. Sharing of model code is also becoming routine, and can be easily achieved through COMSES, Github or similar. The sharing of data in a long-term archive may require more investigation. As a starting project COVID-19 provides an ideal opportunity for setting up such a model inter-comparison project – multiple groups already have running examples, and a shared set of outputs and experiments should be straightforward to agree on. (There will also be huge amounts of data, for example on policy across countries, from which to devise effective comparisons. This also suggests novel research designs. Can a model fitted on one country predict what happened in another for example?) This would potentially form a basis for forward looking experiments designed to assist with possible future pandemic problems (including the need for novel forms of data about things like dynamic contact), and a basis on which to build further features into the existing disease-focussed modelling, such as the effects of economic, social and psychological issues.

5 Additional Challenges for ABMs of Social Phenomena

Nobody supposes that modelling social phenomena is going to have the same set of challenges that climate change models face. Some of the differences include:

- *The availability of good data*. Social science is bedevilled by a paucity of the right kind of data. Although an increasing amount of relevant data is being produced, there are commercial, ethical and data protection barriers to accessing it and the data rarely concerns the same set of actors or events.

- *The understanding of micro-level behaviour*. Whilst the micro-level understanding of our atmosphere is very well established, that of the behaviour of the most important actors (humans) is not. However, it may be that better data (or more attention to the appropriate use of qualitative methods) might partially substitute for a generic behavioural model of decision-making.
- *Agreement upon the goals of modelling*. Although there will always be considerable variation in terms of what is wanted from a model of any particular social phenomena, a common core of agreed objectives will help focus any comparison and give confidence via ensembles of projections. Although the MIPs and Covid Forecast Hub are focussed on prediction, it may be that empirical explanation may be more important in other areas. An additional challenge here is a "higher level" agreement that whatever aims a model has, these should be subject to scientific assessment. Not all ABM should be empirical necessarily, but empirical ABM do have a clear methodology which some other approaches seem to lack.
- *The available resources*. ABM projects tend to be add-ons to larger endeavours and based around short-term grant funding. The funding for big ABM projects is yet to be established, not having the equivalent of weather forecasting to piggyback on. In fact, there may be a Catch-22 here. In order to show what ABM can achieve (and render it suitable for significant ongoing funding by interested policy makers) it may have to self-organise the sort of project that would normally require significant ongoing funding.
- *Persistence of modelling teams/projects*. ABM tends to be quite short-term with each project developing a new model for a new project. This has made it hard to keep good modelling teams together.
- *Deep uncertainty.* Whilst the set of possible factors and processes involved in a climate change model are well established, the basis on which particular mechanisms should feature in a model of any particular social phenomena is currently unclear (but see [4.] for an preliminary attempt to investigate this issue). To take a relevant example from COVID-19, network approaches are clearly important to many social interactions (visiting the houses of others) but not to all (whether you catch COVID-19 in a shop). While the "compartmental" approach is good at representing the transitions of individuals through disease states it fails to allow for the fundamentally social (not physiological) role of agency and policy in changing some of these transitions. Unfortunately, deep disagreements about the assumptions which models require are often bundled with the agendas of different research methods and modelling approaches and the arbitrary exclusion of mechanisms on theoretical or technical grounds can lead to sharp divergences in outcome. Whilst uncertainty in known mechanisms can be quantified, assessing the impact of deep uncertainty and the risk of this kind of mis-specification is much harder.
- *The sensitivity of the political context.* Even in the case of Climate Change, where the assumptions made are relatively well understood and can be justified on objective bases, the modelling exercise and its outcomes can be politically contested. In other areas, where the representation of people's behaviour might be key to model outcomes, this challenge will need even more attention [1.].

However, some of these problems were solved in the case of Climate Change as a result of the CMIP exercise itself and the reports it ultimately resulted in. Over time the development of the models also allowed for a broadening and updating of modelling goals, starting from a relatively narrow initial set of experiments. Ensuring the persistence of individual modelling teams is easier in the context of an internationally recognised comparison project, because resources may be easier to obtain, and there is a consistent central focus. The modelling projects became longer-term as individual researchers could establish a career just doing climate change modelling and the importance of this work increasingly recognised as it became more obviously successful. An ABM modelling comparison project might help solve some of these problems as the importance of its work is established but it would require us to stop telling people that ABM is important and show them it is.

6 Towards an Initial Proposal

Clearly, there are a number of things that could increase the rigour, and hence the reliability, of agent-based modelling. These include better (standardised and more comprehensive) documentation of different aspects of models [12., 13., 14.], free access to the source code of simulations [22., 26.], the organised reproduction of important models [5., 15], being clearer about how data is used [25.] as well as the purpose of a model [9.], and a “reproduction first” system [15.]. However, as argued above, we also need coordinated, large-scale, systematic and persistent model comparison projects. In this section we sketch what this might look like.

The topic chosen for this project should be something where there: (a) is potentially enough public interest to justify the effort, (b) there are a number of models with a similar purpose in mind being developed. At the current stage, this suggests dynamic models of COVID-19 spread, although this may not be a topic of interest for much longer: Whether COVID-19 is a "short term" issue is debatable given the current course of the virus, long-covid, vaccine escape, evolution of new variants and the high level if infection world-wide (and arguably still requires careful long-term modelling). However, the level of current interest among policy makers does not alleviate the responsibility to make models that are capable of representing events that might become policy relevant: it has been suggested for years, for example, that a pandemic was coming, (see e.g. [29]) but when it appeared, little was ready to go by way of models that could deal with the social or economic aspects (and indeed this still seems to be the case, even after 2 years of pandemic). Policy makers continue to show little interest in addressing issues that could cause further global crises (global poverty, the biodiversity crisis, chemical pollution, planetary boundaries in general, or whether exponential economic growth is possible or sensible on a finite planet), but this makes it more, rather than less urgent, to have credible models well-tested models available. Even without considering global crises, there are other possibilities including: transport models (where people go and who they meet) or criminological models (where and when crimes happen). Whichever ensemble of models is chosen, these models should be compared using a core of standards:

- The same start and end dates (but not necessarily the same temporal granularity)
- Covering the same set of regions or cases
- Using the same population data (though possibly enhanced with extra data and maybe scaled population sizes)
- With the same initial conditions in terms of the population
- Outputting a core of agreed measures (but maybe others as well)
- Checked against their agreement with a core set of cases (with agreed data sets)
- Reported in a standard format (though with a discussion section for further/other observations)
- Well documented and with code that is open access
- Run a credible of times with different random seeds

Any modeller/team that had a suitable model and was willing to adhere to the rules would be welcome to participate (commercial, government or academic) and these teams would collectively decide the rules and development of the exercise (along with writing any reports on the resulting comparisons). Other interested stakeholder groups could be involved including professional/academic associations, NGOs and government departments but in a consultative role providing wider critique. It is important that the framework and reports from the exercise be independent or any particular interest or authority.

7 Conclusion

We call upon those who think ABMs have the potential to usefully inform policy decisions to work together, in order that the transparency and rigour of our modelling matches our ambition. Whilst model comparison exercises of various kinds are important for any simulation work, particular care needs to be taken when the outcomes can affect people's lives. *Let us get our act together!*

Acknowledgements

This paper is an expanded version of [3.].

References

1. Aodha, L. & Edmonds, B. Some pitfalls to beware when applying models to issues of policy relevance. In Edmonds, B. & Meyer, R. (eds.) Simulating Social Complexity - a handbook, 2nd edition. Springer, 801-822 (2017). https://doi.org/10.1007/978-3-319-66948-9_29
2. Axtell, R., Axelrod, R., Epstein, J. M., & Cohen, M. D. Aligning simulation models: A case study and results. Computational & Mathematical Organization Theory, 1(2), 123-141 (1996). https://link.springer.com/article/10.1007%2FBF01299065
3. Bithell, M. and Edmonds, B. The Systematic Comparison of Agent-Based Policy Models - It's time we got our act together! Review of Artificial Societies and Social Simulation, 11th May 2021 (2020). https://rofasss.org/2021/05/11/SystComp/
4. Chattoe-Brown, Edmund 'Why Questions Like "Do Networks Matter?" Matter to Methodology: How Agent-Based Modelling Makes It Possible to Answer Them', International Journal of Social Research Methodology (online first 2020). https://doi.org/10.1080/13645579.2020.1801602
5. Chattoe-Brown, E., Gilbert, N., Robertson, D. A., & Watts, C. J. Reproduction as a Means of Evaluating Policy Models: A Case Study of a COVID-19 Simulation. medRxiv 01.29.21250743 (2021); doi: https://doi.org/10.1101/2021.01.29.21250743
6. Climate model inter-comparison project 6, https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6 (*last accessed 19th May 2021*)
7. Covid Forecast Hub, https://covid19forecasthub.org (*last accessed 19th May 2021*)
8. Edmonds, B., & Hales, D. Replication, replication and replication: Some hard lessons from model alignment. Journal of Artificial Societies and Social Simulation, 6(4), 11 (2003). http://jasss.soc.surrey.ac.uk/6/4/11.html
9. Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H. and Squazzoni, F. Different Modelling Purposes' Journal of Artificial Societies and Social Simulation 22(3), 6 (2019). http://jasss.soc.surrey.ac.uk/22/3/6.html>. doi:10.18564/jasss.3993
10. Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E.. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9(5), 1937–1958 (2016). https://doi.org/10.5194/gmd-9-1937-2016

11. Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., Fiorino, M., Gleckler, P. J., Hnilo, J. J., Marlais, S. M., Phillips, T. J., Potter, G. L., Santer, B. D., Sperber, K. R., Taylor, K. E., & Williams, D. N. An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I). In Bulletin of the American Meteorological Society, 80(1), 29–55 (1999). https://doi.org/10.1175/1520-0477(1999)080<0029:AOOTRO>2.0.CO;2
12. Grimm, V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, Goss-Custard J, Grand T, Heinz S K, Huse G, Huth A, Jepsen J U, Jørgensen C, Mooij W M, Müller B, Pe'er G, Piou C, Railsback S F, Robbins A M, Robbins M M, Rossmanith E, Rüger N, Strand E, Souissi S, Stillman R A, Vabø R, Visser U and DeAngelis D L. A standard protocol for describing individual-based and agent-based models. Ecological Modelling 198 (1-2), 115-126 (2006). http://doi.org/10.1016/j.ecolmodel.2006.04.023
13. Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S., ... & Railsback, S. F. Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. Ecological modelling, 280, 129-139 (2014). https://doi.org/10.1016/j.ecolmodel.2014.01.018
14. Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J,, Johnston, A.S.A., Milles, A., Nabe-Nielsen, J., Polhill, J. G., Radchuk, V., Rohwäder, M-S., Stillman, R. A., Thiele, J. C. and Ayllón, D. The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism. Journal of Artificial Societies and Social Simulation 23(2), 7 (2020). <http://jasss.soc.surrey.ac.uk/23/2/7.html>. doi: 10.18564/jasss.4259
15. Hales, D. Vision for a more rigorous “replication first” modelling journal. Review of Artificial Societies and Social Simulation, 5th November 2018. https://rofasss.org/2018/11/05/dh/
16. Hales, D., Rouchier, J., & Edmonds, B. Model-to-model analysis. Journal of Artificial Societies and Social Simulation, 6(4), 5 (2003). http://jasss.soc.surrey.ac.uk/6/4/5.html
17. Jones, B.C., DeBruine, L.M., Flake, J.K. et al. To which world regions does the valence–dominance model of social perception apply?. Nat Hum Behav 5, 159–169 (2021). https://doi.org/10.1038/s41562-020-01007-2
18. Lafuerza L.F., Dyson L., Edmonds B., & McKane A.J. Staged Models for Interdisciplinary Research. PLoS ONE, 11(6): e0157261 (2016). DOI:10.1371/journal.pone.0157261
19. Maxwell, S. E., Lau, M. Y., & Howard, G. S.. Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? American Psychologist, 70(6), 487–498 (2015). https://doi.org/10.1037/a0039400
20. Moshontz, H. + 85 others. The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network, 1(4) 501-515 (2018). https://doi.org/10.1177/2515245918797607
21. Nosek, B. A., & Errington, T. M.. What is replication?. PLoS biology, 18(3), e3000691, (2020). https://doi.org/10.1371/journal.pbio.3000691
22. Polhill, J. G., & Edmonds, B. Open access for social simulation. Journal of Artificial Societies and Social Simulation, 10(3), 10 (2007). http://jasss.soc.surrey.ac.uk/10/3/10.html
23. Psychological Science Accelerator, https://psysciacc.org/ (*last accessed 19th May 2021*)
24. Riolo, R. L., Cohen, M. D., & Axelrod, R. Evolution of cooperation without reciprocity. Nature, 414(6862), 441-443 (2001). https://doi.org/10.1038/35106555

25. Siebers, Peer-Olaf, Achter, Sebastian, Palaretti Bernardo, Cristiane, Borit, Melania and Chattoe-Brown, Edmund. First Steps Towards RAT: A Protocol for Documenting Data Use in the Agent-Based Modeling Process (Extended Abstract), in Ahrweiler, Petra and Neumann, Martin (eds.) Advances in Social Simulation: ESSA 2019, Springer Proceedings in Complexity (Cham: Springer), pp. 257-261 (2021). https://doi.org/10.1007/978-3-030-61503-1_24
26. Squazzoni, F., Polhill, J. G., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., ... & Gilbert, N.. Computational Models That Matter During a Global Pandemic Outbreak: A Call to Action. Journal of Artificial Societies and Social Simulation, 23(2), 10 (2020). http://jasss.soc.surrey.ac.uk/23/2/10.html, DOI: 10.18564/jasss.4298
27. Tittensor, D. P., Eddy, T. D., Lotze, H. K., Galbraith, E. D., Cheung, W., Barange, M., Blanchard, J. L., Bopp, L., Bryndum-Buchholz, A., Büchner, M., Bulman, C., Carozza, D. A., Christensen, V., Coll, M., Dunne, J. P., Fernandes, J. A., Fulton, E. A., Hobday, A. J., Huber, V., … Walker, N. D.. A protocol for the intercomparison of marine fishery and ecosystem models: Fish-MIP v1.0. Geoscientific Model Development, 11(4), 1421–1442 (2018). https://doi.org/10.5194/gmd-11-1421-2018
28. Wei, Y., Liu, S., Huntzinger, D. N., Michalak, A. M., Viovy, N., Post, W. M., Schwalm, C. R., Schaefer, K., Jacobson, A. R., Lu, C., Tian, H., Ricciuto, D. M., Cook, R. B., Mao, J., & Shi, X.. The north american carbon program multi-scale synthesis and terrestrial model intercomparison project - Part 2: Environmental driver data. Geoscientific Model Development, 7(6), 2875–2893 (2014). https://doi.org/10.5194/gmd-7-2875-2014
29. Patterson, Michael M.. "The Coming Influenza Pandemic: Lessons From the Past for the Future"*Journal of Osteopathic Medicine*, vol. 105, no. 11, 2005, pp. 498-500. https://doi.org/10.7556/jom_2005_11.0001