10

SPECIAL TOPIC

Modeling for policy:

Challenges for technology assessment from new prognostic methods

Modellierung für die Politik:

Herausforderungen für die Technikfolgenabschätzung durch neue prognostische Methoden

Edited by Andreas Kaminski, Gabriele Gramelsberger, Dirk Scheer

https://doi.org/10.14512/tatup.32.1.10

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom

INTRODUCTION

Modeling for policy and technology assessment: Challenges from computer-based simulations and artificial intelligence

Andreas Kaminski*[1], Gabriele Gramelsberger[2], Dirk Scheer[3]

Abstract • Modeling for policy has become an integral part of policy making and technology assessment. This became particularly evident to the general public when, during the COVID-19 pandemic, forecasts of infection dynamics based on computer simulations were used to evaluate and justify policy containment measures. Computer models are also playing an increasing role in technology assessment (TA). Computer simulations are used to explore possible futures related to specific technologies, for example, in the area of energy systems analysis. Artificial intelligence (AI) models are also becoming increasingly important. The results is a mix of methods where computer simulations and machine learning converge, posing particular challenges and opening up new research questions. This Special topic brings together case studies from different fields to explore the current state of computational models in general and AI methods in particular for policy and TA.

***Modellierung für Politik und Technologiebewertung**: Herausforderungen durch computergestützte Simulationen und künstliche Intelligenz*

Zusammenfassung • *Modellierung für die Politik ist zu einem integralen Bestandteil der Politikarbeit und der Technologiebewertung geworden. Dies wurde der breiten Öffentlichkeit besonders deutlich, als während der COVID-19-Pandemie Prognosen über die Infektionsdynamik auf Basis von Computersimulationen zur Bewertung und Begründung politischer Maßnahmen zur Eindämmung herangezogen wurden. Computermodelle spielen auch in der Technikfolgenabschätzung (TA) eine zunehmende Rolle. Mithilfe von Computersimulationen werden technologiegebundene Zukunftsperspektiven erkundet, beispielsweise im Bereich der Energiesystemanalyse. Auch Modelle der künstlichen Intelligenz (KI) werden immer wichtiger. Das Ergebnis ist ein Methodenmix, bei dem Computersimulationen und maschinelles Lernen zusammentreffen, was besondere Herausforderungen mit sich bringt und neue Forschungsfragen eröffnet. Dieses TATuP Special topic bringt Fallstudien aus verschiedenen Bereichen zusammen, um den aktuellen Stand von Computermodellen im Allgemeinen und KI-Methoden im Besonderen für Politik und TA zu untersuchen.*

Keywords • *computer-based modeling, technology assessment, artificial intelligence, decision-making, prognostic methods*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

* *Corresponding author: andreas.kaminski@tu-darmstadt.de*
[1] *Department of Philosophy, Technical University of Darmstadt, Darmstadt, DE*
[2] *Human Technology Center, RWTH Aachen University, Aachen, DE*
[3] *Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, DE*

https://doi.org/10.14512/tatup.32.1.11
Received: 01. 02. 2023; accepted: 08. 02. 2023;
published online: 23. 03. 2023 (editorial peer review)

Introduction

The use of models in science has long been a subject of reflection. The philosophy of science has intensively studied the role models play in science. Questions concerning the relationship between model, theory, and experiment, or the potential changes they bring to scientific practice have been addressed here (Morgan and Morrison 1999; Gelfert 2016). The study of modeling for policy, on the other hand, has only recently become

a more intense subject of interdisciplinary reflection (Scheer 2013; Mansnerus 2015). Here, too, the focus is on processes of change, particularly with regard to questions like: How does decision-making change when it is based on, or at least guided by, models?

The history of modeling for policy (as distinct from the study of this history) is older. Models to guide policy questions existed in 18th century demography for example, such as equilibrium models between population size and resources (Fuhrmann 2001). This period also saw the emergence of the first models that play a role in medical policy contexts (see below). These models are primarily conceptual and mathematical in nature. In addition, there are material models, especially in technology development, but also, for example, in hydrology, such as the Large-Scale Mississippi River Basin Model, which is about 200 hectares in size and was used in the mid-20th century to predict or to reproduce past floods (Cheramie 2011). The decline of this (expensive) model in the 1970s coincides with the rise of current modeling approaches, namely computer models.

Computer models are currently mainly computer simulations and artificial intelligence (AI) methods. They are used, e.g., to study the transformation of energy markets, the effects of climate change and possible political-economic strategies for its mitigation, or urban policy design options. Mathematical techniques developed since the 18th century, such as probability theory or numerics, play an important role. However, it would be false to see computer models simply as a continuation of pen-and-paper calculations. The computer is not just a fast calculator. Rather, it changes modeling and the relationship between people and models. For information technology brings its own demands on how and what can be calculated – and in the process it changes the relationship between people (Heymann et al. 2017) causal processes) and thus promoted understanding, today there is an attempt to reduce the opacity of models (Kaminski 2018).

However, it would be equally wrong to write the history of modeling for policy primarily as a history of technology. Not only do models help to shape policy; on the contrary, it is precisely forms of political rationality such as the great planning euphoria or the 'culture of prediction' (Heymann et al. 2017) that promote modeling for policy. The connection between modeling and policy is an act of mutual adaptation, of promotion or inhibition, as well as it is a demand for understanding and a threat of misunderstanding.

The prognostic turn

Modeling for policy and technology assessment has a history that dates back to emerging field of futures studies in the 1950s and 1960s. In particular, Operations Research methods, advanced by Olaf Helmer at the RAND Corporation for 'Long range forecasting', defined that the "future is no longer viewed as unique, unforeseeable, and inevitable; there are, instead, a multitude of possible futures, with associated probabilities that can be estimated and, to some extent, manipulated" (Helmer 1967, p. 2). Operations research, a military term originally used to describe groups of researchers working on large-scale projects such as radar development and surveillance, evolved in the late 1940s into a mathematical method of decision support based on control theory, game theory, linear optimization, and graph theory. Philip M. Morse, who is considered the founder of operations research, wrote in 1945: „Its object is, by the analysis of past operations, to find means of improving the execution of future operations" (Morse and Kimball 1951, p. 5). To this end, Morse had nearly 100 analysts at his disposal in the Operations Research Group founded by the U. S. Navy in 1942, who dealt with questions such as the optimal size of ship convoys or the tactics of air attacks. The success of operations research resulted from the changing situation of warfare under technological conditions. However, the management of changing situations applied not only to military but also to industrial and social conditions in general, which led to a spread of mathematical analysis and planning methods into policy processes (Greenberger et al. 1976; Seefried 2014). Policy problems "differ from operational problems in that unambiguous, rigorous representations of the problems are very difficult to construct." (Kraemer and King 1986, p. 501). Thus, Systems Analysis was developed for the analysis of complex systems under environmental conditions, again at the RAND Corporation. Here, the future came into play less as an optimization of the past than as a statistical uncertainty about the unknown effects of actions, or about unpredictable developments. This growing arsenal of analysis and prediction methods was used for establishing the field of quantitative policy analysis. Big modeling for policy projects were established in New York (Miller et al. 1988) as well as in the Netherlands for water management and storm-surge barriers (Goemans and Visser 1987).

Modeling for quantitative policy analysis became prominent when Limits to Growth was published in 1972 using Jay W. For-

In the analysis of complex systems, the future came into play less as an optimization of the past than as a statistical uncertainty about the unknown effects of actions, or about unpredictable developments.

 https://doi.org/10.14512/tatup.32.1.11

rester's World Dynamics approach developed at the Massachusetts Institute of Technology in Boston (Forrester 1971; Meadows et al. 1972). Commissioned by the Club of Rome, the study calculated the end of mankind by 2100, if no drastic policy measures would be taken. This caused worldwide media interest, which not only made 'modeling for policy' a publicly recognized topic, but anchored the 'prognostic turn' into science, society, and politics. Beside societal studies on the impact of economic growth and exploitation of nature, the increasing environmental problems of pollution, acid rains, and climate change quickly dominated the emerging 'cultures of prediction' (Heymann et al. 2017). Climate modeling, energy modeling and integrated assessment modeling (van Beek et al. 2020), gained importance in the 1970s as well as the international conferences on the increasing concerns of the Earth's condition such as the UN Conference on the Human Environment in Stockholm in 1972. In particular, the awareness of and the political dealing with the ozone hole became a role-model for the emerging global framing of model-based policy analysis and technology assessment. The 'discovery' of the Antarctic ozone hole in 1985 and the assumption that chlorofluorocarbons (CFC) are the sources of the reduction of ozone led already in 1987 to the adoption and in 1989 to the ratification of the Montreal protocol. Ironically, the ozone hole wasn't seen either in the satellite data nor in the model-based simulations before 1985. In-situ measurements, on the other hand, showed a dramatic decrease in ozone (Farman et al. 1985), which called for immediate action – although the "response of policymakers to Farman and colleagues' paper was initially cool." (Solomon 2019, p. 46) However, later models with improved atmospheric chemistry showed that a growing ozone hole would not allow humanity to enjoy full sun for more than ten minutes in 2020 without significantly increasing the likelihood of skin cancer. Although the CFC-assumption wasn't fully verified in 1987, the precautionary principle was firstly applied in the Montreal protocol on a global level. Thus, the protocol not only "prevented the ozone layer from collapsing" and gave us today's freedom of movement under the open sky, it became the "signature success story for global environmental policy" (Solomon 2019, p. 47) and shaped the framing of global climate policy until today.

Model and policy: working on their alignment

If the ozone hole was the paradigm for the global challenge of the 1980s for model-based policy analysis and technology assessment, the COVID-19 pandemic is the global challenge of today. However, the beginnings of deciding health policy issues based on models can be traced back to at least the 18th century: When the number of people dying from smallpox reached a peak, variolation became known as an immunization method in England. This procedure was not without its dangers, insofar as it could itself be fatal or could contribute to the transmission of the disease. Daniele Bernoulli addressed this question on the basis of a probability calculus he had developed, in which he calculated the probabilistic life expectancy of a model population with and without variolation. In 1760, he initially published only the results (1766 then the calculations too) in which he strongly recommended variolation. Bernoulli's contemporary Jean-Baptiste Alembert strongly criticized Bernoulli's approach. An intense debate arose around this early Model for Policy (Colombo and Diamanti 2015; Dietz and Heesterbeek 2002).

Toy models, which characterized the transmission of measles such as the mixing of gas molecules in a tube, and compartment models followed at the beginning of the 20th century (Mansnerus 2015, pp. 12). Then, over the course of the 20th century, models and eventually computer models were developed that examined measles infection or smallpox vaccination strategies, for example, to prepare policy recommendations (Grüne-Yanoff 2017).

More recently, COVID-19 simulations have even come to the attention of a broader audience. In Germany, the Federal Institute for Population Research (BiB) had begun to predict the load of intensive care units in Germany on the basis of a computer simulation. This project exemplifies the work required on the alignment. Indeed, the BiB soon discovered that the model developed to inform policymakers about the predicted situation in intensive care units in German hospitals, in order to derive a basis for COVID-19 measures, was becoming too computationally intensive. Thereby, a start was made with the Federal High-Performance Computing Center (HLRS) at the University of Stuttgart. HLRS had previously hosted several major modeling for policy research projects, such as HiDALGO, in which one of the pilots was to predict the escape movements of people in crisis situations facing war or natural disasters. It quickly turned out that the code developed did not run efficiently on the computers there. As Ralf Schneider noted in a lecture given in the seminar 'Modeling for Policy' at RWTH Aachen on 11.05.2021, a re-implementation of the model became necessary. This revealed a first form of necessary alignment: The way of thinking and coding of the researchers at BiB and the simulation scientists at HLRS had to be aligned under time pressure. The following observations go back to discussions we had with the simulation scientists there: About 20 % of the German population was then represented in the model, and the model was fed with actual data from 401 local counties (Klüsener et al. 2020). The results were forwarded weekly to the RKI and the Federal Ministry of Health. Here, a second alignment became necessary. This concerned the alignment of scientists and politicians. The question arose of whether to work with scenarios and, if so, in what way. The concern on the part of the simulation scientists was in particular that the results would be interpreted in the sense of a weather forecast.

Since the project showed how time-consuming (in a situation that required fast information) this alignment is already among scientists, a follow-up project was created: Computational Immediate Response Center for Emergencies. From our point of view, this project aims to facilitate and stabilize the required epistemic and policy alignment.

Added-value and limitations of computer-based models for policy

We will now take a step further in the reconstruction of how and to what extent this alignment can be achieved: "Policy-making in pluralistic societies is bound to principles of forward-thinking, decision-orientation and evidence-based rationales." (Scharpf 1973) Policies result from a process in which problems to be solved are identified, policy objectives and solutions are then formulated and finally decided by the legislator. Policy interventions are thus key aspects of a decision-based understanding of policy-making (Scheer et al. 2021, p. 7). Computer based models and (lately) AI are to a great extent compatible with these three policy-making features. Decisions about prognostic futures have to be made despite all the complexity of the sociotechnical system, possible path dependencies and uncertainties as well as non-knowledge about (un-)intended economic, ecological and societal consequences of these decisions. The genesis of scientific system, orientation and action knowledge for possible futures plays a central role as an input provider for boundary conditions and impact chains and is confronted with analytical and methodological challenges. However, there are several features of computer-based models and AI that are highly compatible with policy-making. Key characteristics of computer simulations can be synthesized into the following specific capabilities (for the following points and considerations see Scheer 2017, pp. 105–107):

- Display cause-impact chains: Simulations show the effects and outcomes of complex and multidimensional cause-impact relations.
- Reduction of complexity: From a system perspective computational modeling reduce, represent and visualize real-world phenomena, interrelations and statuses.
- Comparison of options: Computer simulations are able to demonstrate and compare several options and courses of action for future developments.
- Intervention effects: With computational modeling the effect and impact of several policy actions, instruments and interventions can be calculated and displayed.
- Formats of results: Simulation results are highly aggregated technical calculations transforming time-depended system states into easily accessible formats of pictures, diagrams and numbers.
- Trial without error: Computer simulations are virtual trial and error operations for finding optimal solutions where the error is not costly and painful.

Computer simulations are science-based instruments for producing knowledge on upcoming future developments. Hence, simulations are an essential addition to the policy impact assessment toolbox and are able to advise policy-makers with relevant information. Using computer simulations, complex real-world systems are reduced to their structural system functions, are replicated in a simplified system 'copy' as a digital twin, and are visible through various visualization techniques. A substantial advantage of simulation is to run system dynamics over time and display various complex system statuses at a specific date where researchers and decision-makers have an interest in. Thus, scientific modeling is a future research and foresight knowledge instrument which may serve as a basis for decisions. The future observing feature of simulations matches perfectly with the forward-looking need of policy-making.

Computed quantitative results in pictures and numbers tend to obscure underlying uncertainties and suggest a level of accuracy which is often not adequate to reality.

Another added value is the comparative character of modeling with relatively easy to do configurations once the principal model is set up. Simulations and scenarios are closely linked in modeling. With slightly changing initial and framework conditions through parameter settings in simulations, modelers are able to compare different scenarios of possible future system developments. With modifications of influencing factors (e.g., parameters) modelers are able to analyze impact and effect of specific (policy) interventions with a trial-and-error method – using a virtual environment without a serious real-world damage. Thus, simulations combine the abilities to run through several alternatives with a clear focus which marks the differences, and the observation of its results and impacts in order to find an optimal solution.

However, computer simulations have their limitations when it comes to policy advice and decision-making. Simulations are often seen as opaque, and thus policy decisions based on simulations are vulnerable and may take center-stage in political dispute over solutions and strategies. The backbone of simulations, that is complexity reduction, comparison of options and policy intervention, are frequently based on oversimplified system functions, starting point assumptions and cause-impact relationships. What is often neglected in simulations are one-time effects and contingencies of human action. On the other side,

 https://doi.org/10.14512/tatup.32.1.11

computed quantitative results in pictures and numbers tend to obscure underlying uncertainties and suggest a level of accuracy which is often not adequate to reality. Against this background, it is not surprising to see that computer simulations are heavily criticized in the policy arena. The main features of simulation critique are a lack of trust in models and modelers, spurious accuracy of simulation results, and inadequacy of the computing process itself which is usually not understandable by the audience.

Model-driven and AI-driven policy analysis and TA

History as well as case studies show that policy analysis is driven by the use of computer-based models and simulations from the very beginning on. However, also technology assessment (TA) is increasingly using modeling and simulation techniques as assessment tools for an anticipatory, "hermeneutic approach" (Grunwald 2022). As policy requirements for technology designs become more demanding – in particular, in terms of sustainability – TA turns from an ad-hoc approach into a prognostic task. Due to the complexity of today's technology designs, prognostic TA "by hand, however, is time-consuming and seems inappropriate" as the case of conceptual aircraft and system research demonstrates (Gradel et al. 2022, p. 281). Therefore, prognostic in-silico TA based on modeling and simulation is required to meet the ambitious political aims of the European Commission's Green Deal (European Commission 2021). "Model-based safety assessment (MBSA) […] uses models to describe the fault behavior of a system. Consequently, safety analyses (e.g., the synthesis of fault trees) can be performed partly automatized with these models." (Gradel et al. 2022, pp. 281–282).

In particular, in Health Technology Assessment (HTA) models have been used to better understand and predict the outcome of policy changes. Again, sustainability – here the UN's Sustainable Development Goals (SDGs) calling for achieving a universal health coverage – is the main driver for the use of prognostic methods (Kingkaew et al. 2022). Interestingly, HTA is also leading in the application of AI methods, although this trend is nascent. "In health care, with the increasing use of information systems and access to large amounts of data, the application of AI tools might facilitate the evidence base of policy decisions. Specifically, in the field of HTA, researchers can rely on health systems data such as administrative claims or electronic health records to generate evidence on health outcomes to support decisions of policy makers and inform patients about the utilization practice, effectiveness, or costs of technologies." (Tachkov et al. 2022, p. 2) While AI technologies are on the forefront of healthcare, for instance for automatic diagnostics, drug development, care robotics, and data management (Davenport and Kalakota 2019), the use AI in healthcare applications still has to be assessed beyond technical performance. In particular, IBM's Watson Oncology failure in 2017 displayed an 'AI chasm' between laboratory conditions and clinical application. Thus, "it becomes clear that regulatory and decision-making organizations as well as HTA agencies are facing unprecedented complexity: evaluating and approving so-called disruptive technologies, especially AI, requires taking several issues into consideration altogether." (Alami et al. 2020, p. 6). A comprehensive TA framework for evaluating technology that uses AI is still lacking.

The contributions in this Special topic

Against the outlined backdrop of the history of model- and AI-based policy analysis and technology assessment this TATuP Special topic 'Modeling for Policy' collects seven papers from scholars of TA, sociology and philosophy of science and technology. We called for contributions that investigate whether and, if so, how decisions change, if they are made on the basis of AI and computer models. Do options for action, evaluations, forecasts or justifications change when policy making decisions are made on the basis of models? In addition, on a second level, to what extent does this change technology assessment, insofar as computer-based models are used to assess technologies? Does it change the courses of action considered in TA? These questions are of interest as AI models and simulations models present a dual challenge for technology assessment.

Firstly, these prognostic methods are used in the object domain of TA. Secondly, TA makes use of these methods itself. In our view, this raises far-reaching epistemic as well as normative questions for TA. This dual challenge concerns, for example, the transparency of TA: the opacity of the models is inherited as a possible opacity of TA. Questions also arise about the robustness of models, especially in novel domains, which then appear as questions about the evaluation of values in TA: is reliability something more important than comprehensibility? Although, the contributions explore different questions and cases, all contributions explore the alignments and frictions, tensions and convergences of models and policies.

In Health Technology Assessment models have been used to better understand and predict the outcome of policy changes.

Anja Bauer and Daniela Fuchs ask in their paper 'Modeling for nano risk assessment and management: The development of

https://doi.org/10.14512/tatup.32.1.11

integrated governance tools and the potential role of technology assessment' for critical reflection these tools from the outside as well as from inside by actively engaging in their development processes. Based on the case of the SUNDS tool both authors show that the tool manifests conceptual shifts from risk to innovation governance.

Lou Therese Brandner and Simon David Hirsbrunner are looking at an entirely different field. Their paper 'Algorithmic fairness in investigative policing: Ethical analysis of machine learning methods for facial recognition' askes fundamental questions about fairness in AI based policing using facial recognition by addressing the AI chasm. Furthermore, they argue that quantitative fairness methods can distract from how discrimination and oppression translate into social phenomena.

Do options for action, evaluations, forecasts or justifications change when policy making decisions are made on the basis of models?

Jens Hälterlein investigates the important case of 'Agent-based modeling and simulation for pandemic management.' He shows that decisions based on these simulations influenced the course of the pandemic and that the use of computer simulations can be understood as a co-production of knowledge about the recent COVID-19 pandemic.

Catharina Landström explores stakeholder involvement in water management in her paper 'Why won't water managers use new scientific computer models? The co-production of a perceived science-practice gap.' She asks, if more stakeholder involvement would lead to an increased uptake of scientific models in water management?

Lilla Horvath, Erich Renz and Christian Rohwer are reflecting on the advantages of 'Combining behavioral insights with artificial intelligence for technology assessment.' As policy decisions concerning technology applications can have far-reaching societal consequences rationality-enhancing procedures are essential. TA will face this challenge.

Titus Udrea, Leo Capari and Anja Bauer examine how models can structure epistemic communities in order to better assess the knowledge claims and evidence politics of computer modeling. Therefore, their paper 'The Politics of Models: Socio-political discourses in modeling of energy transition and transnational trade policies' compares two modeling communities, energy transition and transnational trade.

Johannes Weyer, Fabian Adelt and Marlon Philipp explore 'Pathways to sustainable mobility. Modeling the impact of policy measures' using the example of the Ruhr region and the mobility of the people living there. Simulation experiments show significant differences in the behavior of actor types and in their response to policy interventions. Thus, modeling can help policymakers when planning and designing measures whose goal is sustainable transformation.

References

Alami, Hassane; Lehoux, Pascale; Auclair, Yannick (2020): Artificial intelligence and health technology assessment. Anticipating a new level of complexity. In: Journal of Medical Internet Research 22 (7), p. e17707. https://doi.org/10.2196/17707

Cheramie, Kristi (2011): The scale of nature. Modeling the Mississippi River. In: Places Journal 133 (4), pp. 724–739. https://doi.org/10.22269/110321

Colombo, Camilla; Diamanti, Mirko (2015): The smallpox vaccine. The dispute between Bernoulli and d'Alembert and the calculus of probabilities. In: Lettera Matematica 2 (4), pp. 185–192. https://doi.org/10.1007/s40329-015-0073-5

Dietz, Klaus; Heesterbeek, Hans (2002): Daniel Bernoulli's epidemiological model revisited. In: Mathematical Biosciences 180 (1–2), pp. 1–21. https://doi.org/10.1016/S0025-5564(02)00122-0

Davenport, Thomas; Kalakota, Ravi (2019): The potential for artificial intelligence in healthcare. In: Future Healthcare Journal 6 (2), pp. 94–98. https://doi.org/10.7861/futurehosp.6-2-94

European Commission (2021): European Green Deal. Commission proposes transformation of EU economy and society to meet climate ambitions. Press Release. Brussels: Press material from the Commission Spokesperson's Service. Available online at https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3541, last accessed on 02. 02. 2023.

Farman, Joseph; Gardiner, Brian; Shanklin, Jon (1985): Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction. In: Nature 315 (6016), pp. 207–210. https://doi.org/10.1038/315207a0

Forrester, Jay (1971): World dynamics. Cambridge, MA: Wright Allen Press.

Fuhrmann, Martin (2001): Die Politik der Volksvermehrung und Menschenveredelung. Der Bevölkerungsdiskurs in der politischen und ökonomischen Theorie der deutschen Aufklärung. In: Aufklärung 13, pp. 243–282.

Gelfert, Axel (2016): How to do science with models. A philosophical primer. Dordrecht: Springer. https://doi.org/10.1007/978-3-319-27954-1

Goemans, Tom; Visser, Tjebbe (1987): The Delta Project. The Netherlands experience with a megaproject for flood protection. In: Technology in Society 9 (1), pp. 97–111. https://doi.org/10.1016/0160-791X(87)90034-0

Gradel, Simon; Aigner, Benedikt; Stumpf, Eike (2022): Model-based safety assessment for conceptual aircraft systems design. In: CEAS Aeronautical Journal 13 (1), pp. 281–294. https://doi.org/10.1007/s13272-021-00562-2

Greenberger, Martin; Crenson, Matthew; Crissey, Brian (1976): Models in the policy process. Public decision making in the computer era. New York: Russell Sage Foundation.

Grüne-Yanoff, Till (2017): Seven problems for massive simulation models. In: Michael Resch, Andreas Kaminski and Petra Gehring (eds.): The science and art of simulation. Berlin: Springer, pp. 85–101. https://doi.org/10.1007/978-3-319-55762-5_7

Grunwald, Armin (2022): Model-based anticipation in technology assessment. The hermeneutic approach for opening up a critical perspective. Paper presented at the 4th International Conference on Anticipation (ANTICIPATION 2022), Tempe, AZ, USA, 16. 11. 2022 to 18. 11. 2022.

Helmer, Olaf (1967): Analysis of the future. The Delphi method. Santa Monica: RAND Corporation.

Heymann, Matthias; Gramelsberger, Gabriele; Mahony Martin (eds.) (2017): Cultures of prediction in atmospheric and climate science. Epistemic and cultural shifts in computer-based modeling and simulation. London: Routledge. https://doi.org/10.4324/9781315406282

Kaminski, Andreas (2018): Der Erfolg der Modellierung und das Ende der Modelle. Epistemische Opazität in der Computersimulation. In: Andreas Brenneis, Oliver Honer, Sina Keesser, Annette Ripper and Silke Vetter-Schultheiß (eds.): Technik – Macht – Raum. Das Topologische Manifest im Kontext interdisziplinärer Studien. Wiesbaden: Springer, pp. 317–333. https://doi.org/10.1007/978-3-658-15154-6_16

Kingkaew, Pritaporn et al. (2022): A model-based study to estimate the health and economic impact of health technology assessment in Thailand. In: International Journal of Technology Assessment in Health Care 38 (1), p. e45. https://doi.org/10.1017/S0266462322000277

Klüsener, Sebastian et al. (2020): Forecasting intensive care unit demand during the COVID-19 pandemic. A spatial age-structured microsimulation model. Preprint, In: medRxiv – the preprint server for health sciences, pp. 1–41.

Kraemer, Kenneth; King, John (1986): OR Practice-computer-based models for policy making. Uses and impacts in the U. S. Federal Government. In: Operations Research 34 (4), pp. 501–512. https://doi.org/10.1287/opre.34.4.501

Mansnerus, Erika (2015): Modelling in public health research. How mathematical techniques keep us healthy. New York: Palgrave Macmillan.

Meadows, Donella; Meadows, Dennis; Randers, Jørgen; Behrens, William (1972): The limits to growth. Washington: Potomac Associates Books.

Miller, Louis; Fisher, Gene; Walker, Warren; Wolf Jr., Charles (1988): Operations research and policy analysis at RAND, 1968–1988. In: OR/MS Today 15 (6), pp. 20–25. https://doi.org/10.7249/N2937

Morgan, Mary; Morrison, Margaret (eds.) (1999): Models as mediators. Perspectives on natural and social science. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511660108

Morse, Philip; Kimball, George (1951): Methods of operations research. New York: Technology Press. https://doi.org/10.1063/1.3067068

Scharpf, Fritz (1973): Planung als politischer Prozess. Aufsätze zur Theorie der planenden Demokratie. Frankfurt a. M.: Suhrkamp.

Scheer, Dirk (2013): Computersimulationen in politischen Entscheidungsprozessen. Zur Politikrelevanz von Simulationswissen am Beispiel der CO_2-Speicherung. Wiesbaden: Springer.

Scheer, Dirk (2017): Between knowledge and action. Conceptualizing scientific simulation and policy-making. In: Michael Resch, Andreas Kaminski and Petra Gehring (eds.): The science and art of simulation I. Exploring, understanding, knowing. Cham: Springer, pp. 103–118. https://doi.org/10.1007/978-3-319-55762-5_8

Scheer, Dirk; Class, Holger; Flemisch, Bernd (2021): Introduction. In: Subsurface environmental modelling between science and policy. Cham: Springer, pp. 1–12. https://doi.org/10.1007/978-3-030-51178-4_1

Seefried, Elke (2014): Steering the future. The emergence of "Western" futures research and its production of expertise, 1950s to early 1970s. In: European Journal of Futures Research 2 (1), pp. 291–12. https://doi.org/10.1007/s40309-013-0029-y

Solomon, Susan (2019): The discovery of the Antarctic ozone hole. In: Nature 575 (7781), pp. 46–47. https://doi.org/10.1038/d41586-019-02837-5

Tachkov, Konstantin et al. (2022): Barriers to use artificial intelligence methodologies in health technology assessment in Central and East European countries. In: Frontiers in Public Health 10, p. 921226. https://doi.org/10.3389/fpubh.2022.921226

van Beek, Lisette; Hajer, Maarten; Pelzer, Peter; van Vuuren, Detlef; Cassen, Christoph (2020): Anticipating futures through models. The rise of Integrated Assessment Modeling in the climate science-policy interface since 1970. In: Global Environmental Change 65, p. 102191. https://doi.org/10.1016/j.gloenvcha.2020.102191

PROF. DR. ANDREAS KAMINSKI
is Professor for Philosophy of Science and Technology at TU Darmstadt since 2022. He was head of the department for Philosophy of Computational Science at the Federal High-Performance Computing Center Stuttgart (HLRS). His research focuses on the connection of science and technology (especially in computational science) and on philosophy of trust and testimony.

PROF. DR. GABRIELE GRAMELSBERGER
is Professor for Theory of Science and Technology since 2017 at the RWTH Aachen University. Her research focus lies on the philosophy of computational sciences. Since 2021 she is Director of the Käte Hamburger Kolleg 'Cultures of Research'.

PD DR. DIRK SCHEER
is Senior Researcher at the Institute for Technology Assessment and Systems Analysis at the Karlsruhe Institute of Technology since 2017. His research focuses on social-science based energy research, technology acceptance research, knowledge transfer and management at the science-policy interface, participation and risk research.

RESEARCH ARTICLE

Modeling for nano risk assessment and management: The development of integrated governance tools and the potential role of technology assessment

Anja Bauer*,[1], Daniela Fuchs[2]

Abstract • In nano risk governance, we observe a trend toward coupling and integrating a variety of computational models into integrated risk governance tools. This article discusses the development and design of such integrated tools as 'nano risk governance imaginaries in the making.' Using an illustrative example, the SUNDS tool, we show how the tool manifests conceptual shifts from risk to innovation governance, a technocratic evidence culture based on the quantification of risks, and an envisioned application in industrial innovation management. This conceptualization runs the risk of narrowing the view of nano risks and cementing the widely lamented democratic deficit in risk governance. We therefore conclude that the development and application of integrated governance tools are highly relevant for technology assessment (TA) and TA should actively engage in their development processes.

***Modelle für Risikobewertung und -management von Nanomaterialien**: Die Entwicklung integrierter Governance-Instrumente und die potenzielle Rolle der Technikfolgenabschätzung*

Zusammenfassung • *In der Nanorisiko-Governance beobachten wir einen Trend zur Kopplung und Integration einer Vielzahl computerbasierter Modelle zu integrierten Governance-Instrumenten. In diesem Beitrag wird die Entwicklung und Gestaltung solcher Instrumente als 'Risiko-Governance-Imaginaries im Entstehen' betrachtet. Anhand eines illustrativen Beispiels, dem SUNDS-Tool, zeigen wir, wie das Tool konzeptionelle Verschiebungen von der Risiko- zur Innovations-Governance, eine technokratische Evidenzkultur, basierend auf der Quantifizierung von Risiken, und eine geplante Anwendung im industriellen Innovationsmanagement manifestiert. Diese Konzipierung birgt die Gefahr einer verengten Betrachtungsweise von Nanorisiken und der Zementierung des weithin beklagten Demokratiedefizits in der Risiko-Governance. Wir folgern daher, dass die Entwicklung und Anwendung integrierter Governance-Instrumente für die Technikfolgenabschätzung (TA) von großer Bedeutung sind und TA sich aktiv in deren Entwicklungsprozesse einbringen sollte*

Keywords • *nanomaterials, risk governance, in silico methods, governance imaginary, technology assessment*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Introduction

Nanomaterials have been recognised as promising since the late 1990s, offering research and innovation opportunities in diverse areas such as energy, medicine, electronics, or food. Early on, these expectations were accompanied by concerns about unintended consequences on human health and the environment. Consequently, nanomaterials have increasingly become the subject of regulatory debates and initiatives in the EU and internationally, encouraging the quest for reliable and efficient risk assessment and management approaches.

* *Corresponding author: anja.bauer@aau.at*
[1] *Department of Science, Technologyand Society Studies, University of Klagenfurt, Klagenfurt, AT*
[2] *Institute of Technology Assessment, Austrian Academy of Sciences, Vienna, AT*

https://doi.org/10.14512/tatup.32.1.18
Received: 31. 08. 2022; revised version accepted: 17. 01. 2023;
published online: 23. 03. 2023 (peer review)

Traditionally, risk assessment has strongly relied on 'in vivo' (i.e., on animals) and 'in vitro' (i.e., on cells) methods. The abundance of nanomaterials, unclear effect mechanisms, and ethical concerns about animal testing have challenged these testing regimes and fostered the interest in 'in silico', i.e. computational methods in scientific and policy communities (Worth et al. 2017). Over the past 15 years, a wide variety of models has been developed for distinct risk assessment aspects such as environmental release, fate and exposure, or the toxicological effects of nanomaterials (Isigonis et al. 2019). Besides, computational models have been developed for risk management approaches like control banding. In recent years, we have observed a growing interest in coupling and integrating these single models into integrated risk governance tools. In the following, we use the shortened terms 'integrated tools' or 'tools', implying that they consist of several models. The term 'tool' is used because it is the term used in the respective community and debate and because it refers to its intended use by non-modelers. Over the past decade, the European Commission has funded a range of projects that aimed at developing and testing such integrated tools, linking a variety of screening and assessment methodologies with management, communication, and monitoring tools (EU NanoSafety Cluster 2022; Isigonis et al. 2019). Developing such integrated tools involves various scientific disciplines, including toxicology, lifecycle assessment, or computer sciences, as well as industry partners and consultancies. We observe that technology assessment (TA) or social sciences are less involved in developing such tools and if they are, their role is often limited to ensuring the integration of user needs. The absence of TA is puzzling

Analytical techniques for assessing nano risks have only rarely been addressed or even critically reflected upon by technology assessment.

given its long and active engagement in nano risk governance debates. TA has brought attention to the safety of nanoparticles for human and environmental health early on, presented the respective state of knowledge and uncertainties, frequently organized public dialogues, advanced risk communication and facilitated respective governance structures. In comparison, the analytical techniques for assessing nano risks (e.g., in toxicology), have only rarely been addressed or even critically reflected upon.

This suggests a continued divide between scientific-technocratic approaches to risk governance, (limiting TA and broader societal perspectives), and reflexive approaches of anticipatory governance, including general awareness of risks and the facilitation of public risk debates and communication (as core to TA). We challenge this division by pointing to the value-laden and political nature of analytical techniques in risk governance (Hartley and Kokotovich 2018). Drawing on an illustrative case, the SUNDS tool (section 2), we discuss the development and design of integrated tools as 'nano risk governance imaginaries in the making' (section 3). Concludingly, we reflect on the potential role of TA in such analytical and technical processes (section 4).

SUNDS as an example of integrated governance tools

Our discussion draws on the EU project Sustainable Nanotechnologies (SUN, 2013–2017) and the resulting integrated governance tool SUNDS (Sustainable Nanotechnologies Project Decision Support System). SUNDS (2022) serves as an illustrative case for integrated tools, i.e. to demonstrate and understand how ideas of risks and their governance are (re)produced in tool development and design, without claiming representativeness for all such projects and tools. We perceive SUNDS as an information-rich case (Patton 1990) because the project had been concluded by the time of our research, provided open access to the resulting tool and its documentation, and has been assessed as a comparatively well-designed tool by stakeholders and scientists (Isigonis et al. 2019). Our case study was informed by a critical reading of respective modeling literature and project documentation and ten semi-structured interviews with seven scientists and model developers, one consultant and one industry stakeholder that have been involved in the project as well as one regulator. The interviews focused on the development process and envisioned application of SUNDS, have been recorded, transcribed, and thematically analysed.

SUN was funded under the 7th Framework Programme with a budget of more than 13 million euro. The consortium consisted of 35 partners, including universities, other public and private research institutes, consultancies, organisations for technology transfer, and companies. The project aimed to assess environmental and health risks of manufactured nanomaterials along their lifecycle and to develop the governance tool SUNDS. Industrial partners were involved in testing the tool in product case studies. Moreover, stakeholders from industry, regulation and the insurance sector were engaged through interviews and workshops to ensure the tool's usability.

SUNDS is a web tool for sustainable manufacturing which conducts risk assessment of manufactured nanomaterials and nano-enabled products and determines risk management strategies. It consists of two tiers: The first low-threshold tier includes models for screening environmental, economic, and societal benefits and environmental, occupational and consumer risks of nanoproducts; the second tier offers models to assess environmental life cycle impacts and economic and social aspects in different modules. Users are "expected to insert test results from in-house-tests and literature or to run exposure and haz-

https://doi.org/10.14512/tatup.32.1.18

ard models connected to the SUNDS tool" (Malsch et al. 2017, p. 466). A decision-support module allows the weighing of alternatives including non-nano options and defining risk management strategies (Subramanian et al. 2016). SUNDS has been adapted in subsequent projects, notably in the Horizon 2020 project 'Performance testing, calibration and implementation of a next generation system-of-systems Risk Governance Framework for nanomaterials' (CaLIBRAte, 2016–2019).

Risk governance imaginaries in the making

In developing integrated tools, key issues of nano governance are raised: What should be assessed and why? How can risks be assessed, and by whom? How should risks and benefits be weighted? Who should use the tool, and how? These are not neutral technical questions, but they entangle scientific paradigms and policy discourses (Demortain 2017). We, therefore, conceive of the SUN project and the SUNDS tool as manifestations of risk governance imaginaries. Transferring the notion of "sociotechnical imaginaries" by Jasanoff and Kim (2015, p. 4) from the level of the nation state to the level of scientific and policy areas, we understand risk governance imaginaries in our context as collectively held visions of the (future) governance of nanomaterial risks. Imaginaries motivate and become materialised in sociotechnical developments, in our case, integrated computational tools for nano risk governance. These tools, in turn, structure how we think about nano risk assessment and management, making the collective vision permanent.

In the following, we trace the risk governance imaginary in the making along three dimensions that were central in the tool development. First, the tool development involves explicit and implicit framing activities, inter alia, concerning the need and purpose for risk governance and the conception of risks and relevant assessment dimensions. Second, the development and use of analytical techniques, including computational models, are interwoven with particular evidential cultures, i.e. ways of producing evidence of risks (Boullier et al. 2019; Böschen 2013). Third, our analysis showed that the envisioned application of the tool strongly guided and challenged the development process.

Conceptual framing: towards innovation governance

Questions of why nano risk assessment and management tools are needed, what should be assessed, and which dimensions and principles should guide the assessment, have been vital for the design of SUNDS. Interviews and project documentation indicate that the SUN project and SUNDS tool are firmly embedded in a unison narrative of environmental and health risks potentially hindering industrial innovations. The purpose of SUNDS is to anticipate those risks to facilitate innovation:

> "[I]magine that nanotechnology is a boat […] that includes all the stakeholders – the innovators producing nanotechnology, the regulators making sure that the risks are assessed, the policymakers that are steering the ship – and the ship is going to the shore of making innovation. But there is the sea of uncertainty [where] you have multiple risks like storms, like icebergs, and then the people in the ship are trying to steer the ship to the shore of innovation by dealing with all the risks […] And [SUNDS] is one of the tools to detect the risks and to help the ship navigate in a way to avoid the risks and to reach innovation and shorten the time of reaching innovation" (I3, scientist).

Such framing includes clear value choices favouring nanotechnology innovations and market development. Innovations should be facilitated as efficiently as possible by detecting risks early on. This framing largely excludes questions regarding the innovations' social desirability or acceptance. Its orientation towards sustainable manufacturing determines the conceptualisation of the tool. For example, instead of nanoparticles, SUNDS assesses manufactured nanomaterials and nano-enabled products to better reflect their use in consumer and industry products. This implies the assessment of risks based on actual exposure of affected groups (e.g., consumers), which, according to an industry partner (I2), allows for more realistic scenarios. In addition, the focus on the life cycle allows for assessing risks from the synthesis of the material to the production, use and disposal of the final product. While lifecycle analysis (LCA) traditionally focuses on environmental indicators, SUNDS also offers socioeconomic assessments. This broadening of the assessment not only indicates a more comprehensive view in the light of sustainability but also supports the innovation agenda:

> "[…] since we were using LCA for the environmental aspects, we were trying to align [with] the LCA for the economic parts and social LCA […]. REACH[1] – they have two modes, when you submit something for authorisation, […], you have to either show that you control the risks well or you show that you cannot control the risks, but then you have to show that this is a really important product for the economy and there is no substitute. And the social benefit of having this is unique, so we have no substitute, so we are going to go there even though there are some risks. […] we wanted to [use] this kind of thinking […], so we are not thinking in direction of: oh, there is a harm, let's take it out, but a little bit *how to push the sustainability profile of your product forward"* (I10, scientist, own emphasis).

As the quote suggests, the widening of the scope towards economic and social aspects serves to weigh (environmental and health) risks against (social and economic) benefits to ease in-

1 Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) is a European Union regulation dating from 2006, addressing the production and use of chemical substances, and their potential impacts on human health and the environment.

https://doi.org/10.14512/tatup.32.1.18

dustrial innovations. Lastly, the project took the first steps in integrating the innovation concept 'safety-by-design' into the tool, which was continued in the follow-up project CaLIBRAte and shifted the focus from assessing the final product to reducing hazards from the start.

In sum, SUNDS reflects ongoing debates about moving from risk to innovation governance (Isigonis 2019). This conceptual shift includes important value decisions, such as the balance between innovation and precaution or the (individual or societal) weighting of the environmental, social and economic dimension. The project consortium discussed whether the tool should reflect preferences of users regarding individual analyses (SUN consortium 2015). However, such discussions were closed in favour of the less controversial approach to assigning equal weight to each module.

Evidential culture: quantifying risks, communicating uncertainties

Different ways and techniques exist to produce evidence of risks. The notion of 'evidential culture' "refers to strategies and criteria that frame the collective validation of knowledge" (Boullier et al. 2019, p. 139); in our case, how risks can and should be assessed, whose expertise is sought and how to deal with uncertainties.

In developing SUNDS, different epistemic traditions and cultures came together, from ecotoxicology and lifecycle assessment to human health risk assessment and computational modeling. An understanding of the tool as science-based served as the unifying basis (SUN consortium 2015) and the tool's development was guided by the premise that risks are predictable, measurable, and calculable, putting the quantification of risks (and benefits) at the core. Because of this quantitative paradigm, SUNDS is highly dependent on a wide variety of data; in turn, a lack of accurate, high-quality, and available data is considered the most limiting factor for modeling (I1, consultant). Therefore, SUN dedicated significant resources to collect, systematise and validate existing data. Moreover, as typical in risk assessment methods, semi-quantitative and qualitative approaches were considered additionally or as an approximation, e.g., by using screening tier one when data are missing (Subramanian et al. 2016).

Another challenging task was the integration of various risks and impacts (I6, scientist). While the project team initially intended for the tool to provide 'one single number', this turned out to be non-desirable for stakeholders:

> "Initially, we wanted to integrate everything in a single score […] But […] one of the key findings of the stakeholder consultations was: 'even if you gave us a single number in the end with respect to the sustainability of the material, what would we do with it? We would rather see a dashboard, seeing this is going well, this is going kind of well, and this is not going well […].' – A single number – how would a stakeholder know how to interpret that? How to improve their product?" (I10, scientist).

Hence, stakeholders considered the inclusion of semi-quantitative or qualitative elements and transparent communication of individual assessments as a valuable feature of integrated tools, while the scientists' and modelers' aspiration to quantify and therewith to gain 'more accurate' results persisted (I10, scientist).

Uncertainties in the modeled results and their communication was a core issue in SUN. On one side, the project aimed to reduce models' uncertainties by gathering experimental evidence (Marcomini and Hristozov 2017). On the other side, transparency and communication of uncertainties were recognised as key to risk management. Accordingly, SUNDS explicates uncertainties in the outputs and provides users with uncertainty analyses, including magnitude and sources of model uncertainty (Isigonis et al. 2019).

In sum, we observe both the reproduction of a technocratic evidential culture of risk assessment based on quantitative methods and the consideration of reflexive elements by strengthening qualitative and semi-quantitative approaches or focusing on risk communication. The quantitative paradigm (in particular of tier two) guides the assessment focus towards those things that can be measured and, in the short run, also those things for which data exist, while other less measurable aspects (risk perceptions or different framings) may be excluded. Moreover, the hegemony of scientific risk assessment expertise is stabilised: The tool development strongly relied on scientific expertise and industry stakeholders, yet broader social scientific expertise or societal perspectives were marginal. Actor groups with potentially critical stances, such as consumer groups, health activists or environmental NGOs, were considered mainly as 'imagined actors', i.e., their perspectives were included as a context factor, yet not directly sought.

Envisioned application: REACH(ing) experts

Tools like SUNDS are developed with a view to their application for specific tasks in risk governance. Thus, their envisioned function and the roles of different groups vis-à-vis the model are crucial. In which situations, by whom, and how should the tool be used?

Though specifications for nanomaterials only came into force after the SUN project had been finalised, REACH served as the central reference point for SUNDS, implying its application in the context of industrial risk assessment and management to comply with current (and future) regulations. While initially, the SUN project aimed to address policymakers, industry and the insurance sector (Malsch et al. 2017), the focus on REACH combined with diverging stakeholder interests narrowed down the main target groups:

> "[…] you try to wield the tool for so many users and at some point, you realise … the insurance sector we could not help much. […] The regulators told us what would be acceptable scientifically and […] submission and stuff, but we ended up *majorly building the tool for industry*.

https://doi.org/10.14512/tatup.32.1.18

> The regulators were on board, but it was just very difficult bridging all of them" (I10, scientist, own emphasis).

Thus, the tool became more exclusionary during its development and eventually was designed for large industries and small and medium enterprises (SMEs) (I3, scientist). Concerning when and how to use the tool, the REACH context suggests a use for risk-benefit analyses. Yet, the tool's design goes beyond this application and facilitates guidance on decision-making in risk management (I3, scientist) and transparent communication. For example, the tool proposes "technological alternatives and risk management measures to reduce risks to acceptable levels" and allows the "comparison of scenarios with and without these measures" (Isigonis et al. 2019, p. 14). Most prominently, it features workplace safety measures like protective gear and technical equipment to be selected for the respective risk scenario. While the extension of the tool beyond regulatory demands, in principle, allows for its broader application in industrial innovation management, it also adds to its complexity, with consequences for its potential to be adopted. Due to the science-based and data-driven nature, combined with the multitude of assessments, the tool ultimately depends on a high level of technical expertise and a large amount of data to be used (I1, consultant).

Interviewees strongly suggest that the tool may not be usable for all intended users; particularly SMEs might be excluded, by design rather than intention, as they often neither have the expertise nor data to appropriately use the tool (SUN consortium 2015, p. 3). Partly, this limitation is accepted for developing a science-based and comprehensive tool. Moreover, this exclusion is partly mitigated by the modular design of the tool, with the semi-quantitative tier one being useable for most stakeholders and the fully-fledged risk assessment of tier two being targeted to experts in large companies (I10, scientist). Moreover, further activities to facilitate the application have been undertaken in follow-up projects. Still, the tension between a high degree of scientific rigour (drawing on data and quantification) and pragmatic usability persists.

In sum, the strong orientation towards the regulatory context restricts the range of intended users of the tool, excluding various stakeholders and non-expert publics from its use, even if it provides open access. Beyond that, we observe a further unintended narrowing down of potential users due to the tool's comprehensive, complex and data-driven design. In the short run, this might result in the non-utilisation of the tool. In the long run, a wider (mandatory) use of modeling tools may imply that some actors (e.g., SMEs) need to adapt their risk assessment and management practices.

Discussion and conclusions

In this article, we sketched the risk governance imaginary that became materialised in the integrated tool SUNDS, characterised by a conceptual focus on innovation management, a technocratic evidential culture based on the quantification of risks, and an envisioned application by industry experts. This imaginary partly reflects ongoing debates on nano risk governance and regulatory contexts, notably REACH with the demand for risk-benefit considerations or the focus on industry. However, the tool also exceeds current regulations when shifting further towards innovation governance and including additional analyses. By incorporating broader expectations and visions about future regulatory and governance needs, the tool may have performative effects on future regulatory regimes by technically allowing specific questions and assessments or by including or excluding actors.

In our discussion, SUNDS served as an illustrative yet not representative example of the design of integrated governance tools. Other tools may include different concepts or technical design choices, yet our argument that a particular imaginary of nano risk governance materialises in computational tools holds. Since the tools are tailored to specific regulatory demands and the concerns of specific groups with specific ideas about risks, there is a danger of narrowing the ways of seeing, debating and assessing risks (Demortain 2017, p. 145). Such tools may disguise value choices in favour of "technological innovation and market development in scientific methods […] of quantifying the risks and benefits of technologies" (Demortain 2017, p. 145). Moreover, technologically-framed rather than socially-framed risk assessment (and governance) tools (McLaren 2018) exclude or marginalise actors such as environmental and health activists. In this way, the respective governance imaginaries imply a risk of closing down nano risk governance and further cementing the widely lamented democracy deficit, which TA has long aimed to counter.

Concludingly, we suggest a role for TA in countering the closing down of nano risks governance imaginaries by engaging in the debates and development processes of analytical techniques such as integrated computational tools. In doing so, TA can draw on its broad repertoire of advancing nano risk debates in other sites (e.g., policy fora or public deliberations). As our analysis has illustrated, the tool development and respective imaginary touch upon a range of issues that have long been of interest for TA, including risk communication, the balance between precaution and innovation, risk perceptions, decisions under conditions of uncertainty, the balancing of environmental, social and economic concerns or the inclusion of wider societal perspectives in risk debates.

First, TA could open up conceptual discussions beyond user preferences by clarifying the broader visions, dominant framings, and values that guide the tool development. This opening could also challenge taken-for-granted assumptions and goals, for example, about innovations, sustainability, or consumer safety. Second, building on its long tradition of fostering participation in technology governance, TA could guide the development of such tools towards more inclusive and democratic activities. Thus, stakeholder inclusion could be widened from users towards broader societal participation that includes lay publics

 https://doi.org/10.14512/tatup.32.1.18

and alternative or counter expertise (for example, environmental or consumer NGOs). Scholars have recently outlined how to best integrate these perspectives into risk assessment and governance (Hartley and Kokotovich 2018; Hartley et al. 2022). Lastly, we see a role for TA in promoting alternative ways in which computational models could be employed in governance. What is underrepresented in the discussions of such tools is their potential to serve as boundary objects to coordinate different actors and their perspectives (Star and Griesemer 1989). This would, for example, mean strengthening discussions around objectives, parameters, values, and scenarios. Such alternative visions of the tools' functions exist, pointing to its use in developing regulations or to use them in international governance (Malsch et al. 2018), yet remain marginalised. TA could help make such perspectives more prominent.

Funding • This article is based on research conducted in the project 'CoMoPA – Computational Modelling for Policy Advice' which received funding from the Austrian Academy of Science Innovation Fund 'Research, Science and Society' [no. IF_2017_13].

Competing interests • The authors declare no competing interests.

References

Böschen, Stefan (2013): Modes of constructing evidence. Sustainable development as social experimentation. The cases of chemical regulations and climate change politics. In: Nature and Culture 8 (1), pp. 74–96. https://doi.org/10.3167/nc.2013.080105

Boullier, Henri; Demortain, David; Zeeman, Maurice (2019): Inventing prediction for regulation. In: Science & Technology Studies 32 (4), pp. 137–157. https://doi.org/10.23987/sts.65062

Demortain, David (2017): Expertise, regulatory science and the evaluation of technology and risk. Introduction to the special issue. In: Minerva 55 (2), pp. 139–159. https://doi.org/10.1007/s11024-017-9325-1

EU NanoSafety Cluster (2022): Strategic direction enhancing synergies. A high profile platform for the coordination of nanosafety research in Europe. Available online at https://www.nanosafetycluster.eu, last accessed on 17. 01. 2023.

Hartley, Sarah; Kokotovich, Adam (2018): Disentangling risk assessment. New roles for experts and publics. In: Brigitte Nerlich, Sarah Hartley, Sujatha Raman and Alexander Smith (eds.): Science and the politics of openness. Here be monsters. Manchester: Manchester University Press, pp. 176–194. https://doi.org/10.7765/9781526106476.00019

Hartley, Sarah; Kokotovich, Adam; McCalman, Caroline (2022): Prescribing engagement in environmental risk assessment for gene drive technology. In: Regulation & Governance. Early View. https://doi.org/10.1111/rego.12452

Isigonis, Panagiotis et al. (2019): Risk governance of nanomaterials. Review of criteria and tools for risk communication, evaluation, and mitigation. In: Nanomaterials 9 (5), pp. 1–26. https://doi.org/10.3390/nano9050696

Jasanoff, Sheila; Kim, Sang-Hyun (eds.) (2015): Dreamscapes of modernity. Sociotechnical imagineries and the fabrication of power. Chicago: Unversity of Chicago Press. https://doi.org/10.7208/chicago/9780226276663.001.0001

Malsch, Ineke et al. (2017): Comparing mental models of prospective users of the sustainable nanotechnology decision support system. In: Environment Systems and Decisions 37, pp. 465–483. https://doi.org/10.1007/s10669-017-9648-3

Malsch, Ineke; Mullins, Martin; Semenzin, Elena; Zabeo, Alex; Hristozov, Danail; Marcomini, Antonio (2018): Decision support for international agreements regulating nanomaterials. In: NanoEthics 12 (1), pp. 39–54. https://doi.org/10.1007/s11569-018-0312-2

Marcomini, Antonio; Hristozov, Danail (2017): Sustainable nanotechnologies (SUN). Final report summary. Available online at https://cordis.europa.eu/project/id/604305/reporting, last accessed on 16. 01. 2023.

McLaren, Duncan (2018): Whose climate and whose ethics? Conceptions of justice in solar geoengineering modelling. In: Energy Research & Social Science 44, pp. 209–221. https://doi.org/10.1016/j.erss.2018.05.021

Patton, Michael (1990): Qualitative evaluation and research methods. In: Research in Nursing & Health 14 (1), pp. 73–74. https://doi.org/10.1002/nur.4770140111

Star, Susan; Griesemer, James (1989): Institutional ecology, 'translations' and boundary objects. Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907–39. In: Social Studies of Science 19 (3), pp. 387–420. https://doi.org/10.1177/030631289019003001

Subramanian, Vrishali et al. (2016): Sustainable nanotechnology decision support system. Bridging risk management, sustainable innovation and risk governance. In: Journal of Nanoparticle Research 18 (4), pp. 1–13. https://doi.org/10.1007/s11051-016-3375-4

SUN consortium (2015): Summary report on SUN user workshop. Final version. Available online at http://www.sun-fp7.eu/wp-content/uploads/2015/02/SUN-user-workshopsummaryfinal.pdf, last accessed on 04. 01. 2023.

SUNDS (2022): Decision support system for risk assessment and management of nano(bio)materials used in consumer products and medical applications. Available online at https://sunds.gd/?request=/sections, last accessed on 04. 01. 2023.

Worth, Andrew et al. (2017): Evaluation of the availability and applicability of computational approaches in the safety assessment of nanomaterials. Final report of the Nanocomput project. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/911484

DR. ANJA BAUER
is Assistant Professor at the Department of Science, Technology and Society Studies at the University of Klagenfurt. She researches and teaches in the areas of environmental, sustainability and technology governance with a special interest in the role of expertise, anticipation, and participation in policy-making.

DANIELA FUCHS
is a Junior Scientist at the Institute of Technology Assessment at the Austrian Academy of Sciences. Her research focus is on the governance of emerging technologies.

https://doi.org/10.14512/tatup.32.1.18

RESEARCH ARTICLE

Algorithmische Fairness in der polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung

Lou Therese Brandner*,[1], Simon David Hirsbrunner[1]

Zusammenfassung • Dieser Beitrag diskutiert Fairness in auf künstlicher Intelligenz (KI) basierenden Verfahren der Polizeiarbeit anhand des Beispiels der Gesichtserkennung. Algorithmische Entscheidungen, die auf gesellschaftlichen Diskriminierungsdynamiken beruhen, können Ungerechtigkeiten (re-)produzieren und automatisieren. KI-Fairness betrifft dabei nicht nur die Erstellung und das Teilen von Datensätzen oder das Training von Modellen, sondern auch die Art des Systemeinsatzes in der Realwelt. Die Quantifizierung von Fairness kann davon ablenken, wie Diskriminierung und Unterdrückung sich konkret als soziale Phänomene niederschlagen. Integrative Ansätze können hier dazu beitragen, durch kontinuierliche interdisziplinäre Kollaboration ethische, rechtliche, soziale und wirtschaftliche Faktoren aktiv in die Technikentwicklung einzubeziehen und die Folgen des Einsatzes ganzheitlicher einzuschätzen.

***Algorithmic fairness in investigative policing**: Ethical analysis of machine learning methods for facial recognition*

Abstract • *This article discusses fairness in artificial intelligence (AI) based policing procedures using facial recognition as an example. Algorithmic decisions based on discriminatory dynamics can (re)produce and automate injustice. AI fairness here concerns not only the creation and sharing of datasets or the training of models but also how systems are deployed in the real world. Quantifying fairness can distract from how discrimination and oppression translate concretely into social phenomena. Integrative approaches can help actively incorporate ethical, legal, social, and economic factors into technology development to more holistically assess the consequences of deployment through continuous interdisciplinary collaboration.*

Keywords • *fairness, policing, algorithmic bias, machine learning*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Einleitung

Der Einsatz von so genannter künstlicher Intelligenz (KI) in der polizeilichen Ermittlungsarbeit verspricht effektivere und kostengünstigere Verbrechensprävention und -aufklärung. KI kann dort ansetzen, wo menschliche Fähigkeiten und Kapazitäten potenziell nicht ausreichen; durch automatisierte, papierlose Arbeitsprozesse sollen sowohl die Produktivität als auch die Objektivität polizeilicher Maßnahmen verbessert werden. Diverse technische Lösungen mit KI-Unterstützung für die polizeiliche Ermittlungsarbeit befinden sich auch in Deutschland bereits im Einsatz, bspw. zur Sichtung und Auswertung von kinderpornographischem Bildmaterial (LKA Niedersachsen 2020).

Während die KI-Branche mit großer Geschwindigkeit wächst, wurde erst 2021 mit dem Artificial Intelligence Act der Europäischen Union (EU) der weltweit erste Gesetzentwurf zur KI-Regulierung vorgelegt. Dieser definiert Echtzeit-Gesichtserkennung im öffentlichen Raum als Hochrisikotechnologie, deren Einsatz nur in besonderen Ausnahmefällen wie der Terroris-

* *Corresponding author: lou.brandner@uni-tuebingen.de*
[1] *Internationales Zentrum für Ethik in den Wissenschaften (IZEW), Universität Tübingen, Tübingen, DE*

https://doi.org/10.14512/tatup.32.1.24
Received: 26. 08. 2022; revised version accepted: 17. 01. 2023;
published online: 23. 03. 2023 (peer review)

musbekämpfung oder der Suche nach vermissten Personen erlaubt sein soll (Europäische Kommission 2021). Im Polizeikontext können fehlerhafte Entscheidungen zu gravierenden Grundrechtsverletzungen wie rechtswidrigen Festnahmen führen (Selbst 2017). Um unerwünschte Nebeneffekte und Risiken für Individuen und Personengruppen zu minimieren, sind an diese hochrisikoreichen Anwendungen über technische Ansprüche hinaus besonders strenge Anforderungen auf dem gesellschaftlichen Level zu stellen. Hochrisiko-KI tangiert Fragen sozialer Gerechtigkeit und die Operationalisierung von Prinzipien wie Fairness, Gleichbehandlung und Nichtdiskriminierung.

Diskriminierung grenzt sich von Ungleichbehandlung dadurch ab, dass sie auf der Zugehörigkeit zu sozial bedeutsamen Gruppen beruht (Lippert-Rasmussen 2013). Das bedeutet, dass Personen aufgrund von Attributen wie Geschlecht, ethnischer Herkunft oder Behinderung ungerechtfertigter, negativer Ungleichbehandlung ausgesetzt sind (Hagendorff 2019). Algorithmische Outputs, die auf ‚Biases' (auf Deutsch Verzerrungen) in KI-Systemen beruhen und sich so auf bestehende gesellschaftliche Diskriminierungsdynamiken stützen, können Ungerechtigkeiten (re-)produzieren und automatisieren (Eubanks 2018). Der Begriff KI-Fairness, auch ‚algorithmische Fairness' genannt, beschreibt Methoden, die durch Verzerrungen in KI-Modellen hervorgerufene soziale Diskriminierung mindern oder ausschließen sollen.

Dieser Beitrag diskutiert Fairness in KI-basierten Verfahren der polizeilichen Ermittlungsarbeit. Im Folgenden wird zunächst kurz der Projektkontext erläutert, aus dem die Autor*innen die Beitragselemente ableiten, bevor verschiedene Quellen von Verzerrungen in KI-Systemen aufgezeigt werden. Daraufhin werden KI-Fairness und ihre Grenzen anhand verschiedener Metriken und deren realweltlichen Auswirkungen problematisiert. Der Beitrag schließt ab mit einer Diskussion zum Bezug zur Technikfolgenabschätzung (TA) und der effektiven Problematisierung von KI-Fairness in interdisziplinären Forschungsprojekten.

Hintergrund

Die Überlegungen in diesem Artikel leiten sich aus ethischen Analysen ab, die im Rahmen zweier interdisziplinärer Projekte im Kontext ziviler Sicherheitsforschung durchgeführt wurden, in denen verschiedene digitale Erkennungs- und Analysetechnologien für die polizeiliche Ermittlungsarbeit entwickelt werden. Die Forschung schließt explizit die Adressierung von Bias und Diskriminierung in Verfahren KI-basierter Gesichtserkennung ein. Die Projekte verfolgen einen integrierten Ansatz der Technikentwicklung (Spindler et al. 2020), bei dem technische Perspektiven mit ethischen, rechtlichen, sozialen und wirtschaftlichen (ELSE) Aspekten in Bezug gesetzt werden. Wie bereits im TA-Kontext diskutiert (Gressel und Orlowski 2019), reichen diese Ansätze über den Anspruch klassischer ethischer Begleitforschung hinaus, da ELSE-Aspekte von Projektbeginn an und über alle Arbeitspakete hinweg einbezogen werden.

Dieser Beitrag hat den Anspruch, KI-Fairness in polizeilichen KI-Systemen über einzelne Projekte hinausgehend zu diskutieren. Die Fokussierung auf die Gesichtserkennung bietet sich jedoch an, da sie als weitverbreitete Anwendung viele ethische Problematiken algorithmischer Polizeiarbeit aufzeigt. Biometrische Erkennungssysteme stellen besondere Herausforderungen an die KI-Modellierung, da individuelle menschliche Charakteristiken erfasst, erkannt und verifiziert werden müssen (Merler et al. 2019). Durch die Entwicklung und Anwendung von KI-basierter Gesichtserkennung ergeben sich auch besondere ethische Problematiken, was sich im bestehenden KI-Regulierungsentwurf (Europäische Kommission 2021) der EU widerspiegelt, der die biometrische Fernidentifizierung als Hochrisikotechnologie einstuft. Verzerrungen und resultierende Diskriminierung werden hier explizit als besondere Risiken genannt (Europäische Kommission 2021, S. 26).

Durch die Entwicklung und Anwendung von KI-basierter Gesichtserkennung ergeben sich auch besondere ethische Problematiken.

Die Technologie wird trotz der dargestellten Risiken in diversen Kontexten von Überwachung und Kontrolle erprobt (Bundespolizei 2018; Jürs 2022; Monroy 2022); auch in von den Autor*innen begleiteten Projekten werden Demonstratoren von polizeilichen Gesichtserkennungssystemen entwickelt, um bspw. Personen in Überwachungsaufnahmen zu identifizieren. Dies wirft ethisch-normative Fragestellungen hinsichtlich Fairness und Biases auf, die im Rahmen der Technikentwicklung mit technischen, polizeilichen, kommerziellen und rechtlichen Partnern diskutiert werden. Von dieser Zusammenarbeit ausgehend bietet sich eine Evaluierung der Diskriminierungsrisiken unter den Vorzeichen der TA an. Entsprechend sollen in diesem Beitrag mögliche Verzerrungen aufgeführt und im Hinblick auf Fairness-Kriterien reflektiert werden.

Verzerrungen in KI-Verfahren der Gesichtserkennung

Techniken angewandter KI können zur Identifizierung von Gesichtern in fotografischen Bilddaten verwendet werden. Vor der Operationalisierung in konkreten Einsatzgebieten werden solche Systeme anhand riesiger Mengen von Beispieldaten trainiert und lernen so, verschiedene Gesichter anhand verschiedener visuel-

https://doi.org/10.14512/tatup.32.1.24

ler Merkmale zu unterscheiden. Dabei kommen sowohl überwachte als auch unüberwachte Ansätze des KI-Trainings zum Einsatz (Anwarul und Dahiya 2020). Aus unterschiedlichen Gründen variiert die Treffsicherheit des Systems oft zwischen verschiedenen zu identifizierenden Merkmalen, was potenziell zur Diskriminierung sozialer Gruppen führen kann. Im Folgenden werden Auslöser und Mechanismen solcher Verzerrungen in polizeilich genutzter Gesichtserkennung diskutiert.

Qualität der Trainingsdaten

Eine bedeutende Quelle für Verzerrungen in KI-Systemen sind Trainingsdaten.[1] Im Rahmen von Trainingsprozessen etablieren KI-Modelle auf Basis von Datensätzen bestimmte Ähnlichkeiten und Unterschiede zwischen Elementen. In Verfahren des unüberwachten Lernens entstammen die Daten bspw. sozialen Medien oder Bilddatenbanken, die Verzerrungen enthalten können. Bestimmte Gruppen, z. B. Frauen, nichtweiße oder ältere Personen, können in Trainingsdatenbanken unterrepräsentiert sein, was dazu führt, dass ihnen gegenüber die algorithmische Ergebnisgenauigkeit sinkt (Berk et al. 2018). Wenn wenig historische Daten existieren oder aus Gründen des Datenschutzes nicht nutzbar sind, können synthetische Trainingsdaten reale Datensätze ersetzen oder ergänzen. Diese fiktiven, realistische Daten, die gezielt für die KI-Entwicklung hergestellt werden, sind grundsätzlich dazu geeignet, das Problem existierender Verzerrungen in historischen Daten zu umgehen, können Minderheiten aber auch fehlrepräsentieren (Bhanot et al. 2021).

Auch sind Qualität und Auswertbarkeit von Foto- und Videoaufnahmen häufig von schwierigen Konditionen hinsichtlich z. B. Perspektive, Belichtung oder Bildauflösung beeinflusst, die allgemein zur geringeren Genauigkeit von KI-Modellen, z. B. bezüglich der korrekten Identifizierung von Straftäter*innen, beitragen können (Anwarul und Dahiya 2020). Verlässlichkeit und Genauigkeit und daher auch die Fairness der Modelle hängen so zu einem erheblichen Teil von der Verfügbarkeit qualitativ hochwertiger Trainingsdaten ab.

Annotation der Trainingsdaten

Auch die Einschreibung sozialer und kultureller Vorannahmen während der Datenannotation kann zu Verzerrungen in Trainingsdatensätzen führen (Selbst 2017). In überwachten KI-Trainingsprozessen markieren Personen bestimmte Charakteristika in Datensätzen, deren Erkennung anschließend von einem Modell erlernt wird. Das Labeling von Gesichtsdatenbanken wird üblicherweise von KI-Entwickler*innen selbst oder von Dienstleister*innen auf Crowdsourcing-Plattformen wie Amazon Mechanical Turk oder Scale durchgeführt. Dabei identifizieren menschliche Akteure übereinstimmende Aufnahmen von Individuen und kennzeichnen Merkmale wie Gesichtszüge, Mimik und Posen. Diskriminierende Stereotype können hier durch unbeabsichtigte Voreingenommenheit oder absichtliche Beeinflussung Eingang finden (Leslie 2020). Wenn bspw. Individuen bestimmter Ethnien häufiger falsch identifiziert werden, können die Modelle dies in höheren Fehlerraten für diese Gruppen widerspiegeln. Bei aktiven KI-Lernverfahren, in welchen Nutzer*innen Verantwortung bei der Anpassung und Optimierung automatisierter Analysesysteme eingeräumt werden, kann Diskriminierung zudem im laufenden Betrieb eingeschleust werden, wenn Einzelne den Lernprozess durch oft unbewusste Vorurteile beeinflussen (Fischer et al. 2022).

Ergebnisgenauigkeit der KI-Verfahren

Ein bekanntes Phänomen in der KI-Gesichtserkennung sind signifikant höhere Fehlerquoten für weibliche[2] und nicht-weiße Gesichter und somit eine schlechtere Ergebnisgenauigkeit. Abseits der Repräsentation in Trainingsdaten können hier auch Faktoren wie Kameraeinstellungen (Roth 2009), Gesichtsmorphologie oder Makeup (Albiero et al. 2022) eine Rolle spielen. Diese Umstände bilden gesellschaftliche Dynamiken ab, die weiße und männliche Personen als den menschlichen Standard positionieren und andere Gruppen als Abweichungen marginalisieren. Bestehende Diskriminierung wird so im KI-Kontext fortgeführt, insbesondere gegenüber von intersektionalen Unterdrückungen betroffenen Gruppen wie Schwarzen Frauen (Buolamwini und Gebru 2018).

Nicht nur Unter-, sondern auch Überrepräsentation kann zu Biases führen, wenn z. B. Schwarze Individuen häufiger polizeilich erfasst und somit häufiger durch Datenbankabgleiche – korrekt wie inkorrekt – identifiziert werden (Bacchini und Lorusso 2019). Da eine häufigere Erfassung nicht zwingend eine tatsächliche höhere Straffälligkeit bedeutet, sondern auch durch diskriminierende Polizeipraktiken wie unverhältnismäßige Überwachung oder Verfolgung bedingt ist (Garvie und Frankle 2016; Selbst 2017), werden so existierende rassistische Dynamiken verstärkt.

Überprüfung algorithmischer Fairness

Erschwerend kommt hinzu, dass der Analyseprozess und die identifizierten Datenmuster im maschinellen Lernen, z. B. bei der Identifizierung eines Individuums, wie beschrieben häufig hochkomplex und für Personen kognitiv kaum interpretierbar sind. Gerade der Beitrag, den die gewählten Trainingsdatensätze auf die Outputs eines solchen Blackbox-Modells haben, ist selten konkret nachvollziehbar. Die Opazität und wahrgenommene Autorität von Computerprozessen können dazu führen, dass Biases unerkannt bleiben, die zusätzlich stabilisiert werden, wenn polizeiliche Ermittler*innen ohne tiefergehendes Vorwissen über KI-Systeme scheinbar neutrale algorithmische Entscheidungen unter dem Einfluss subjektiver Vorannahmen interpretieren (Helm und Hagendorff 2021).

1 Der Begriff der Trainingsdaten schließt hier Trainingsdaten (Vorlagen zum Erlernen von Mustern), Validierungsdaten (Optimierung des Modells) und Testdaten (Überprüfung des Modellverhaltens) mit ein.

2 Gängige KI-Modelle kategorisieren biologisches Geschlecht binär als männlich–weiblich und ziehen komplexere soziale Genderkonstruktionen nicht in Betracht. Da dieser Beitrag sich mit der Praxis der Programme befasst, wird die dichotome Einteilung übernommen.

https://doi.org/10.14512/tatup.32.1.24

Lösungsansätze der KI-Fairness

Die Europäische Kommission definiert Fairness im KI-Kontext als den Schutz vor (algorithmischer) Verzerrung, Diskriminierung und Stigmatisierung von Personen und Gruppen (Europäische Kommission 2018). Diese Problematiken werden auf verschiedenen Wegen von Fairness-Ansätzen adressiert. Dort, wo hohe Fehlerquoten durch Unterrepräsentation in Trainingsdaten bedingt sind, können Datensätze mit höherer Diversität hinsichtlich Hautfarbe und Geschlecht wie z. B. Diversity in Faces (Merler et al. 2019) oder Pilot Parliaments Benchmark (Buolamwini und Gebru 2018) Abhilfe schaffen. Bereits formulierte Anforderungen an fairness-sensible Daten (Le Quy et al. 2022) bieten sich zum Vergleich verwendeter Datensätze an. Doch Diversität in Datensätzen ist zwar eine Grundlage, aber kein alleiniger Garant für Fairness.

Statistische Methoden der KI-Fairness sollen durch algorithmische Verzerrungen in KI-Modellen hervorgerufene soziale Diskriminierung mindern oder ausschließen. Gängige informatische Definitionen unterscheiden dabei insbesondere zwischen individueller Fairness und Gruppenfairness. Erstere soll sicherstellen, dass statistische Messungen der Ergebnisse für Individuen mit denselben Merkmalen gleich sind. Bei Verfahren der Gruppenfairness werden Ergebnisse eines Modells so angeglichen, dass sie für verschiedene vordefinierte Gruppen von Datensubjekten mit geschützten Merkmalen ähnlich oder gleich sind (Mahoney et al. 2020, S. vii).

Es existiert eine Vielzahl statistischer Methoden, mit denen sich KI-Fairness evaluieren lässt (Mehrabi et al. 2021). Die Identifizierung des geeigneten Verfahrens in einem konkreten Fall operationalisiert über technische Fragestellungen hinaus verschiedene Vorstellungen von Fairness und Gerechtigkeit. Angenommen, eine KI selektiert erfolgsversprechende Kandidat*innen für ein Informatikstudium: Bei der Optimierung gegenüber dem Attribut binäres Geschlecht auf ‚demographic parity' soll die Verteilung sozialer Gruppen in einer Gesamtbevölkerung abgebildet werden, sodass jeweils die Hälfte der Studienplätze an Frauen und an Männer vergeben wird, unabhängig von der durchschnittlichen Qualität der jeweiligen Bewerbungen (Quotengerechtigkeit). Optimiert auf die Metrik der ‚equalized odds' hingegen, zielt das System darauf, dass Bewerber*innen mit ähnlichen Qualifikationen gleiche Chancen haben, während die letztendliche Geschlechterquote im Studiengang gleichgültig ist (Chancengerechtigkeit).

KI-Verfahren zur Fairness in der Gesichtserkennung werden gerne auf Metriken der Chancengerechtigkeit mit gleichen ‚true-positive'-Raten für geschützte und ungeschützte Gruppen trainiert. Es wird dabei angenommen, dass eine Erkennung durch das System die favorable Option darstellt, was jedoch nicht immer der sozialen Realität entspricht. In der polizeilichen Ermittlungsarbeit bedeuten ‚false positives' in einer Gesichtserkennungs-Software, dass Unbeteiligte als Verdächtige fehlidentifiziert werden; die Folgen von falsch-positiven Ergebnissen sind hier besonders schwerwiegend bis hin zu unbegründeten Festnahmen und Gefährdungen der körperlichen Unversehrtheit. ‚False negatives' bedeuten, dass Täter*innen nicht identifiziert werden und dadurch der Strafverfolgung entgehen. Während der Suche nach einem vermissten Kind wiederum spielen ‚false positives' eine weniger große Rolle, da sie direkt nach der Fehlentscheidung des Systems durch menschliche Akteure überprüfbar sind und daher keine Folgen für das fälschlich identifizierte Kind haben. Ein ‚false negative' resultiert dagegen in einer Nicht-Identifizierung des vermissten Kindes und hat damit potenziell schwerwiegende Nachteile für Betroffene und Ermittlungsbehörden, deren Suche durch den Fehler des Systems verlängert wird oder sogar erfolglos bleibt.

Niedrige Fehlerraten sind in der algorithmischen Polizeiarbeit daher von hoher Priorität, was insbesondere für sozial benachteiligte Gruppen relevant ist, da für diese wie dargelegt Fehlerraten häufig signifikant höher sind. Eine künstliche Angleichung der Raten verringert dabei zwangsweise die Treffgenauigkeit des technischen Systems (Kleinberg et al. 2016) und Fairnessformeln lassen sich oft nicht parallel optimieren, da sie über gemeinsame Variablen verbunden sind (Ruf und Detyniecki 2021). Das Ausklammern geschützter Merkmale (‚fairness through unawareness') (Gajane und Pechenizkiy 2018) kann keine allgemeine Lösung darstellen, da es die Nichterkennung von Biases zur Folge haben kann. So ließe sich beim Ausschluss des Attributs Geschlecht in der Gesichtserkennung eine erhöhte Fehlerrate gegenüber Frauen schlechter feststellen und daher nicht beheben. Zudem können nichtgeschützte Attribute als Stellvertreter für geschützte Merkmale fungieren. Z. B. hat sich gezeigt, dass die an sich unproblematische Variable ‚Postleitzahl' unter Umständen auf das geschützte Merkmal ‚Ethnie' (‚race') schließen lässt (Datta et al. 2017).

Jede Quantifizierung und Abstrahierung von Fairness kann davon ablenken, wie KI-Systeme in strukturellen Diskriminierungsdynamiken verankert sind.

Fairness tangiert auch die Art des Systemeinsatzes in der Realwelt. Angenommen, ein KI-gestütztes System für Gesichtserkennung wird in einem Stadtteil mit mehrheitlich Schwarzer Bevölkerung installiert: Auch wenn das eingesetzte KI-System gleiche Fehlerraten für alle Ethnien ausgibt, stellen sich hier Fragen der Diskriminierung durch KI – eben deshalb, weil eine Schwarze Nachbarschaft von einem besonders präzisen (sogar algorithmisch ‚fairen') System überwacht wird, während andere Stadtteile unbeobachtet bleiben. Jede versuchte Quanti-

https://doi.org/10.14512/tatup.32.1.24

fizierung und Abstrahierung von Fairness kann davon ablenken, wie KI-Systeme in strukturellen Diskriminierungsdynamiken verankert sind (John-Mathews et al. 2022) und wie diese sich konkret im Leben von unterprivilegierten Personen niederschlagen (Birhane et al. 2022). Das tangiert auch andere Voraussetzungen für Fairness wie Transparenz und Erklärbarkeit, da Betroffene der Entscheidungen von Hochrisiko-Systemen z. B. wissen können sollten, auf welche Metrik hin optimiert wurde und welche Ergebnisse bei den verwendeten Verfahren erreicht wurden.

Diskussion und Ausblick

Wie sich im Kontext der hier diskutierten polizeilichen Gesichtserkennung zeigt, ist statistisch verstandene KI-Fairness alleine nicht geeignet, um Nicht-Diskriminierung im realen Betrieb sicherzustellen. Aus ethischer Perspektive darf der Blick nicht von historischen und strukturellen Machtdynamiken abrücken, da eine eng gefasste Auffassung von KI-Fairness zu sehr auf bloße Daten anstatt auf realweltliche Konsequenzen gerichtet sein kann (John-Mathews et al. 2022). Normative Überlegungen sind hier von großer Bedeutung, um die Folgen der Anwendung von Fairnesskriterien und -metriken kontextspezifisch einzuschätzen. Ob und wie ein ‚fairer' Einsatz von Gesichtserkennungssystemen und anderer polizeilicher KI möglich ist und wie die Risiken dieser Anwendungen minimiert werden können, kann nicht allgemeingültig beantwortet werden; Fragen wie diese müssen im Kontext der jeweiligen Szenarien ganzheitlich und aus transdisziplinärer Sicht von den beteiligten Stakeholdern verhandelt werden. Das bekannte Collingridge-Dilemma, demzufolge in frühen Entwicklungsstadien noch Einfluss auf Technologien genommen werden kann, aber noch Unsicherheit über die tatsächliche Nutzung herrscht, während in späten Stadien das Gegenteil gilt, ist im Zusammenhang mit sich häufig noch in der Forschungsphase befindenden KI eine besondere Herausforderung der TA (Humm et al. 2021).

Integrierte Ansätze der Technikentwicklung können dieses Dilemma und verwandte Problematiken effektiv adressieren, da technische Perspektiven von Anfang an und durchgehend mit ELSE-Aspekten in Bezug gesetzt werden. Die für die TA notwendige kontinuierliche und ganzheitliche Beobachtung der Technikentwicklung und Beurteilung einzelner Einsatzformen (Albrecht und Kellermann 2020) werden so ermöglicht. Im Rahmen der Projektarbeit der Autor*innen zeigt sich, dass dieses Vorgehen dazu geeignet ist, normative Probleme wie algorithmische Diskriminierung frühestmöglich zu thematisieren; in den Polizeiprojekten wurde das ethische Teilprojekt bereits zu Anfang in die Formulierung technischer Spezifikationen der Vorhaben zur Gesichtserkennung involviert. So konnten die beschriebenen Quellen von Diskriminierung sowie vorhandene Lösungsansätze und ihre Grenzen kontinuierlich mit technischen, polizeilichen und rechtlichen Partnern diskutiert und neu ausgehandelt werden.

Funding • This article is based on the work in the research projects PEGASUS and VIKING, funded by the German Federal Ministry of Education and Research (BMBF) as part of the federal government's program "Research for Civil Security".

Competing interests • The authors declare no competing interests.

Literatur

Albiero, Vítor; Zhang, Kai; King, Michael; Bowyer, Kevin (2022): Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. In: Transactions on Information Forensics and Security 17, S. 127–137. https://doi.org/10.1109/TIFS.2021.3135750

Albrecht, Thorben; Kellermann, Christian (2020): Künstliche Intelligenz und die Zukunft der digitalen Arbeitsgesellschaft. Konturen einer ganzheitlichen Technikfolgenabschätzung. Working Paper Forschungsförderung, Nr. 200. Düsseldorf: Hans-Böckler-Stiftung.

Anwarul, Shahina; Dahiya, Susheela (2020): A comprehensive review on face recognition methods and factors affecting facial recognition accuracy. In: Pradeep Singh, Arpan Kar, Yashwant Singh, Maheshkumar Kolekar und Sudeep Tanwar (Hg.): Proceedings of ICRIC 2019. Cham: Springer, S. 495–514. https://doi.org/10.1007/978-3-030-29407-6_36

Bacchini, Fabio; Lorusso, Ludovica (2019): Race, again. How face recognition technology reinforces racial discrimination. In: Journal of Information, Communication and Ethics in Society 17 (3), S. 321–335. https://doi.org/10.1108/jices-05-2018-0050

Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Kearns, Michael; Roth, Aaron (2018): Fairness in criminal justice risk assessments. The state of the art. In: Sociological Methods & Research 50 (1), S. 3–44. https://doi.org/10.1177/0049124118782533

Bhanot, Karan; Qi, Miao; Erickson, John; Guyon, Isabelle; Bennett, Kristin (2021): The problem of fairness in synthetic healthcare data. In: Entropy 23 (9), S. 1–21. https://doi.org/10.3390/e23091165

Birhane, Abeba et al. (2022): The forgotten margins of AI ethics. In: FACCT'22. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, S. 948–958. https://doi.org/10.1145/3531146.3533157

Bundespolizei (2018): Projekt zur Gesichtserkennung erfolgreich. Testergebnisse veröffentlicht. Online verfügbar unter www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011_abschlussbericht_gesichtserkennung.html, zuletzt geprüft am 17. 01. 2023.

Buolamwini, Joy; Gebru, Timnit (2018): Gender shades. Intersectional accuracy disparities in commercial gender classification. In: Proceedings of Machine Learning Research 81, S. 1–15. Online verfügbar unter https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf, zuletzt geprüft am 17. 01. 2023.

Datta, Anupam; Fredrikson, Matthew; Ko, Gihyuk; Mardziel, Piotr; Sen, Shayak (2017): Use privacy in data-driven systems. Theory and experiments with machine learnt programs. In: Proceedings of the 2017 ACM SIGSAC Conference, S. 1193–1210. https://doi.org/10.1145/3133956.3134097

Eubanks, Virginia (2018): Automating inequality. How high-tech tools profile, police, and punish the poor. New York: St. Martin's Press. https://doi.org/10.5204/lthj.v1i0.1386

Europäische Kommission (2021): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. 2021/0106 (COD). Brüssel: Europäische Kommission.

Europäische Kommission (2018): Ethik-Leitlinien für eine vertrauenswürdige KI. Brüssel: Europäische Kommission. https://doi.org/10.1007/978-3-663-09857-7_27

Fischer, Maximilian; Hirsbrunner, Simon; Jentner, Wolfgang; Miller, Matthias; Keim, Daniel; Helm, Paula (2022): Promoting ethical awareness in communication analysis. Investigating potentials and limits of visual analytics for intelligence applications. In: FACCT'22. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, S. 877–889. https://doi.org/10.1145/3531146.3533151

Gajane, Pratik; Pechenizkiy, Mykola (2018): On formalizing fairness in prediction with machine learning. In: arxiv.org. https://doi.org/10.48550/ARXIV.1710.03184

Garvie, Clare; Frankle, Jonathan (2016): Facial-recognition software might have a racial bias problem. In: The Atlantic, 07. 04. 2017. Online verfügbar unter www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/, zuletzt geprüft am 17. 01. 2023.

Gressel, Céline; Orlowski, Alexander (2019): Integrierte Technikentwicklung. Herausforderungen, Umsetzungsweisen und Zukunftsimpulse. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 28 (2), S. 71–72. https://doi.org/10.14512/tatup.28.2.s71

Hagendorff, Thilo (2019): Maschinelles Lernen und Diskriminierung. Probleme und Lösungsansätze. In: Österreichische Zeitschrift für Soziologie 44 (S1), S. 53–66. https://doi.org/10.1007/s11614-019-00347-2

Helm, Paula; Hagendorff, Thilo (2021): Beyond the prediction paradigm. Challenges for AI in the struggle against organized crime. In: Law and Contemporary Problems 84 (3), S. 1–17. https://scholarship.law.duke.edu/lcp/vol84/iss3/2

Humm, Bernhard; Lingner, Stephan; Schmidt, Jan; Wendland, Karsten (2021): KI-Systeme. Aktuelle Trends und Entwicklungen aus Perspektive der Technikfolgenabschätzung. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 30 (3), S. 11–16. https://doi.org/10.14512/tatup.30.3.11

John-Mathews, Jean-Marie; Cardon, Dominique; Balagué, Christine (2022): From reality to world. A critical perspective on AI fairness. In: Journal of Business Ethics 178, S. 945–959. https://doi.org/10.1007/s10551-022-05055-8

Jürs, Martin (2022): Hamburg Airport. Lufthansa setzt auf Gesichtserkennung. In: fvw Travel Talk, 29. 04. 2022. Online verfügbar unter www.fvw.de/touristik/verkehr/hamburg-airport-lufthansa-setzt-auf-gesichtserkennung-225741, zuletzt geprüft am 17. 01. 2023.

Kleinberg, Jon; Mullainathan, Sendhil; Raghavan, Manish (2016): Inherent trade-offs in the fair determination of risk scores. In: arxiv.org. https://doi.org/10.48550/arXiv.2203.05051

Le Quy, Tai; Roy, Arjun; Iosifidis, Vasileios; Zhang, Wenbin; Ntoutsi, Eirini (2022): A survey on datasets for fairness-aware machine learning. In: WIREs Data Mining and Knowledge Discovery 12 (3), S. 1–59. https://doi.org/10.1002/widm.1452

Leslie, David (2020): Understanding bias in facial recognition technologies. An explainer. In: SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3705658

Lippert-Rasmussen, Kasper (2013): Born free and equal? A philosophical inquiry into the nature of discrimination. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199796113.001.0001

LKA Niedersachsen (2020): Künstliche Intelligenz. LKA Niedersachsen stellt Software zur Bekämpfung von Kinderpornografie bundesweit zur Verfügung. Online verfügbar unter www.lka.polizei-nds.de/a/presse/pressemeldungen/kuenstliche-intelligenz-lka-niedersachsen-stellt-software-zur-bekaempfung-von-kinderpornografie-bundesweit-zur-verfuegung-114750.html, zuletzt geprüft am 17. 01. 2023.

Mahoney, Trisha; Varshney, Kush; Hind, Michael (2020): AI fairness. Sebastopol: O'Reilly Media. Online verfügbar unter https://krvarshney.github.io/pubs/MahoneyVH2020.pdf, zuletzt geprüft am 17. 01. 2023.

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina; Galstyan, Aram (2021): A survey on bias and fairness in machine learning. In: ACM Computing Surveys 54 (6), S. 1–35. https://doi.org/10.1145/3457607

Merler, Michele; Ratha, Nalini; Feris, Rogerio; Smith, John (2019): Diversity in faces. In: arxiv.org. https://doi.org/10.48550/arXiv.1901.10436

Monroy, Matthias (2022): DNA, Gesichtsbilder und Fingerabdrücke. Biometrische BKA-Systeme enthalten Datenblätter zu zehn Millionen Personen. In: Netzpolitik.org, 09. 03. 2022. Online verfügbar unter https://netzpolitik.org/2022/dna-gesichtsbilder-und-fingerabdruecke-biometrische-bka-systeme-enthalten-datenblaetter-zu-zehn-millionen-personen, zuletzt geprüft am 17. 01. 2023.

Roth, Lorna (2009): Looking at Shirley, the ultimate norm. Colour balance, image technologies, and cognitive equity. In: Canadian Journal of Communication 34 (1), S. 111–136. https://doi.org/10.22230/cjc.2009v34n1a2196

Ruf, Boris; Detyniecki, Marcin (2021): Towards the right kind of fairness in AI. In: arxiv.org. https://doi.org/10.48550/arXiv.2102.08453

Selbst, Andrew (2017): Disparate impact in big data policing. In: Georgia Law Review 52, S. 109–195. http://dx.doi.org/10.2139/ssrn.2819182

Spindler, Mone; Booz, Sophia; Gieseler, Helya; Runschke, Sebastian; Wydra, Sven; Zinsmaier, Judith (2020): How to achieve integration? Methodological concepts and challenges for the integration of ethical, legal, social and economic aspects into technological development. In: Bruno Gransche und Arne Manzeschke (Hg.): Das geteilte Ganze. Horizonte Integrierter Forschung für künftige Mensch-Technik-Verhältnisse. Wiesbaden: Springer, S. 213–240. https://doi.org/10.1007/978-3-658-26342-3

DR. LOU THERESE BRANDNER
ist wissenschaftliche Mitarbeiterin am Internationalen Zentrum für Ethik in den Wissenschaften der Universität Tübingen. Sie hat Soziologie an der Universität von Amsterdam studiert und an der Universität La Sapienza promoviert. Sie forscht zu algorithmischer Überwachung, digitalem Kapitalismus und räumlichen Fragen.

DR. SIMON DAVID HIRSBRUNNER
leitet am IZEW Projekte zu den Themen KI-Ethik, Datenethik, sowie algorithmischer Polizeiarbeit. Er ist Sozial- und Medienwissenschaftler und forschte bisher im Bereich Human-Centered Computing an der Freien Universität Berlin, sowie bei der Wikimedia, am Potsdam Institut für Klimafolgenforschung, der Universität Potsdam und der Universität Siegen.

https://doi.org/10.14512/tatup.32.1.24

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom

RESEARCH ARTICLE

Agentenbasierte Modellierung und Simulation im Pandemiemanagement

Jens Hälterlein*,[1]

Zusammenfassung • Mathematische Modelle und Computersimulationen spielen im Rahmen der COVID-19-Pandemie eine entscheidende Rolle für das Wissen über die mögliche Entwicklung des Infektionsgeschehens und für entsprechende politische Entscheidungen. Der Beitrag stellt Ergebnisse aus einer ethnografischen Studie über ein staatlich finanziertes Forschungs- und Entwicklungsprojekt vor, das sich mit der agentenbasierten Modellierung und Simulation (ABMS) im Kontext des Pandemiemanagements befasst. Ausgehend von der Annahme, dass der Einsatz von Computersimulationen im Pandemiemanagement nicht nur Mittel zum Zweck für politische bzw. epidemiologische Ziele ist, sondern wesentlich mitbestimmt, welche Ziele und Strategien als politisch legitim erscheinen, rekonstruiert der Beitrag, wie in der ABMS und spezifisch im beforschten Projekt Erkenntnisse über die Pandemie generiert und für die Entscheidungsfindung zugänglich gemacht werden.

Agent-based modeling and simulation for pandemic management

Abstract • *Mathematical models and computer simulations play a crucial role in the context of the COVID-19 crisis for knowledge about the possible course of the pandemic and for appropriate policy decisions. The paper presents results from an ethnographic study of a government-funded R & D project dealing with agent-based modeling and simulation (ABMS) in the context of pandemic management. Based on the assumption that the use of computer simulations in pandemic management is not only a means to an end for political or epidemiological goals but also plays a significant role in determining which goals and strategies appear politically legitimate, the paper reconstructs how insights into the pandemic are generated in ABMS and specifically in the researched project and made accessible for decision-making.*

* *Corresponding author: jens.haelterlein@upb.de*

[1] *Institut für Medienwissenschaften, Universität Paderborn, Paderborn, DE*

https://doi.org/10.14512/tatup.32.1.30
Received: 26. 08. 2022; revised version accepted: 17. 01. 2022;
published online: 23. 03. 2023 (peer review)

Keywords • *agent-based modeling, simulation, pandemic management*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Aktuelle Fallzahlen, zeitliche Verlaufskurven und animierte Grafiken des Infektionsgeschehens spielen nicht nur in der medialen Kommunikation zur COVID-19-Pandemie eine zentrale Rolle.[1] Auch politische Entscheidungsträger*innen begründen ihr Handeln in der Regel mit dem Verweis auf ein Wissen, das sich aus dem Monitoring von Neuinfektionen und die Analyse dieser Daten speist. Eine besondere Rolle beim Management der Pandemie nehmen mathematische Modellierungen und Computersimulationen von Infektionsdynamiken ein.

So hatte eine Modellierung, anhand derer die Verbreitung von SARS-CoV-2 in der britischen Bevölkerung in unterschiedlichen Szenarien simuliert wurde, entscheidenden Einfluss auf das Handeln der britischen Regierung. Zu Beginn der Pandemie setzte diese zunächst auf die Strategie der ‚Herdenimmunität'. In Anbetracht eines simulierten Szenarios, bei dem das Gesundheitssystem Großbritanniens durch SARS-CoV-2 Infektionen überlastet werden würde und bis zu 500.000 Todesfälle auf Grund von COVID-19 Infektionen auftreten könnten, entschied sich die britische Regierung jedoch zu einschneidenden Maßnahmen. Für die USA wurde auf Basis des gleichen Modells ein Szenario mit bis zu 2,2 Millionen Todesfällen simuliert. Mit diesen Zahlen konfrontiert, entschied sich das Weiße Haus umgehend Maßnahmen zu ergreifen und erließ u. a. Regeln für ein ‚social distancing' (Adam 2020).

Auch mit Blick auf Deutschland lässt sich ein entscheidender Einfluss von Modellierungen auf politische Entscheidungsprozesse konstatieren.

1 Einzelne Passagen dieses Beitrags wurden bereits in Hälterlein 2020 veröffentlicht.

In einer Stellungnahme der Deutschen Gesellschaft für Epidemiologie vom 18. März 2020 (Deutsche Gesellschaft für Epidemiologie 2020) wurden, basierend auf den Ergebnissen einer Computersimulation, Handlungsempfehlungen in Richtung der Entscheidungsträger auf Bund- und Länderebene ausgesprochen. Mit Blick auf die geschätzten Kapazitätsgrenzen des deutschen Gesundheitssystems wurde mit Nachdruck empfohlen, bereits bestehende Maßnahmen (Verbot von Großveranstaltungen und Schulschließungen) durch weitere Beschränkungen sozialer Kontakte zu ergänzen. Wenige Tage später beschlossen die Bundesregierung und die Länder mit Verweis auf diese Empfehlungen ein umfangreiches Maßnahmenpaket in dessen Kern allgemeine Kontaktbeschränkungen standen.

Für Deutschland lässt sich ein entscheidender Einfluss von Modellierungen auf politische Entscheidungsprozesse konstatieren.

Bereits vor ihrem Erfolg im Zuge der COVID-19 Krise wurden jedoch die Validität der Modelle und die Aussagekraft der auf diesen Modellen basierenden Computersimulationen in Frage gestellt. Es wurde insbesondere bemängelt, dass Modelle in der Regel nicht berücksichtigen, dass Individuen ihr Verhalten in Epidemien/Pandemien anpassen und beispielsweise Großveranstaltungen meiden oder freiwillig Hygienemaßnahmen ergreifen. Diese Anpassungen haben zur Folge, dass sich Infektionsdynamiken ändern, was wiederum dazu führe, dass modellbasierte Vorhersagen, die individuelles Verhalten nicht berücksichtigen, nicht zutreffen (Epstein et al. 2008). Einer der Gründe für diese folgenschwere Abstraktion ist, dass auf mathematischen Gleichungen basierende Modelle schlichtweg von einer homogenen Bevölkerung ausgehen, da sich eine differenzierte Darstellung in ungleich komplexeren Gleichungen niederschlagen müsste, was die Anwendung und das Verständnis der Modelle deutlich erschweren würde (Frias-Martinez et al. 2011). Als vielversprechende Alternative zu diesen Makro-Modellierungen werden nun immer häufiger agentenbasierte Modelle und Simulationen (ABMS) genannt, da diese eine Modellierung individuellen Verhaltens und daher eine deutlich differenziertere und insofern auch realistischere Simulation von Szenarien ermöglichen würden (Lorig et al. 2021). Aus der Perspektive der Technikfolgenabschätzung (TA) gilt es, dieses Versprechen kritisch zu prüfen.

Im Folgenden werde ich Ergebnisse aus einer ethnografischen Studie über ein staatlich finanziertes Forschungs- und Entwicklungsprojekt vorstellen, das sich mit der agentenbasierten Modellierung von Infektionsdynamiken im Kontext des Pandemiemanagements befasst. In dem Projekt wurde ein Demonstrator eines auf verteilter künstlicher Intelligenz (KI) basierenden Systems entwickelt, das Kommunen bei der Bewältigung von Krisenlagen unterstützt, indem auf Basis der Auswertung von ‚Smart-City'-Daten ein Lagebild erstellt wird und mögliche Auswirkungen von lokalen, nicht-pharmazeutischen Maßnahmen simuliert und visualisiert werden. Im Rahmen der ethnografischen Studie wurden Interviews mit am Projekt beteiligten Akteuren geführt, die aus dem Projekt entstandenen Publikationen analysiert und eine teilnehmende Beobachtung bei der öffentlichen Präsentation der Projektergebnisse durchgeführt.

Ausgehend von der Annahme, dass der Einsatz von Computersimulationen im Pandemiemanagement nicht nur Mittel zum Zweck für politische bzw. epidemiologische Ziele ist, sondern wesentlich mitbestimmt, welche Ziele und Strategien als politisch legitim erscheinen, möchte ich rekonstruieren, wie in der ABMS und spezifisch im beforschten Projekt Erkenntnisse über die Pandemie generiert und wie diese Erkenntnisse für die Entscheidungsfindung zugänglich gemacht werden können. Dabei wird eine Reihe von Fragen adressiert, die für die TA von hoher Relevanz sind: Welche Aussagekraft haben die Ergebnisse der agentenbasierten Simulation von Szenarien und wo liegen deren Grenzen? Wie sollten diese Ergebnisse interpretiert werden, wenn sie einerseits auf fehlerhaften Daten und unrealistischen Annahmen über das Verhalten der individuellen Agenten basieren könnten und andererseits der stochastische Charakter des Outputs statistischer, d. h. nicht-deterministischer Modelle berücksichtigt wird? Welche Gefahren bestehen, wenn die Ergebnisse von ABMS nicht ausreichend kritisch reflektiert werden?

Was leisten Modelle und Simulationen in einer Pandemie?

Zunächst gilt es allerdings zu klären, worin der Zweck von Modellierungen und Computersimulationen im Kontext des Pandemiemanagements besteht. Modelle und Computersimulationen ermöglichen das Erstellen unterschiedlicher Szenarien von möglichen Zuständen komplexer Systeme. Epidemiologische Computersimulationen lassen sich demnach als virtuelle Laboratorien begreifen, in denen Wissenschaftler*innen Erkenntnisse über den möglichen zukünftigen Verlauf eines Infektionsgeschehens produzieren. Diese Erkenntnisse sind natürlich auch für politische Entscheidungsträger*innen nützlich: Wenngleich viele, empirisch nicht belegte oder nicht belegbare Annahmen sowie Schätzungen numerischer Werte (Reproduktionszahl, Inkubationszeit, Erkrankungsdauer, Sterblichkeitsrate etc.) in eine epidemiologische Simulation einfließen, ermöglichen die generierten Szenarien eine Kontingenzbewältigung im Umgang mit

https://doi.org/10.14512/tatup.32.1.30

Ungewissheit. Durch sie wird es möglich, potenzielle Auswirkungen von Epidemien zu antizipieren und vorbereitende Maßnahmen für deren Eintreten zu ergreifen. Darüber hinaus ermöglichen sie das Testen von pharmazeutischen und insbesondere nicht-pharmazeutischen Maßnahmen im Hinblick auf deren Effektivität bei der Bekämpfung einer Pandemie. Ein solches experimentelles Durchspielen von Handlungsoptionen unter Laborbedingungen vermag die Unsicherheit zu reduzieren, die mit der Unvorhersehbarkeit von Handlungseffekten in der politischen Praxis einhergeht (Opitz 2017). Epidemiologische Computersimulationen können also politische Entscheidungsprozesse unterstützen, indem sie die möglichen Konsequenzen von Entscheidungen im Rahmen dieser Entscheidungen antizipierbar machen.

Aus diesem Grund plädiere ich dafür, den Einsatz von epidemiologischen Computersimulationen als Prozess algorithmischer Entscheidungsfindung, dem sogenannten ‚algorithmic decision-making' (ADM), zu verstehen. Der Grad an Automatisierung variiert innerhalb des ADM. Im Falle von epidemiologischen Computersimulationen handelt es sich um Entscheidungsassistenzsysteme, die Entscheidungen nicht selbst treffen, sondern unterstützen. Die Entscheidungshoheit verbleibt damit bei den Anwender*innen des Systems (‚human in the loop').

Eine der interviewten Personen verdeutlicht den Doppelcharakter von epidemiologischen Computersimulationen als Erkenntnis- und Entscheidungsinstrument anhand des Dashboards, das im Projekt entwickelt wurde und als Mensch-Maschine-Schnittstelle fungiert, über die Benutzer*innen mit der ABMS in Interaktion treten können. Sie beschreibt das Dashboard analog zu einem Flugzeugcockpit:

> „Und zwar ist die Idee dabei letztendlich, dass […], wenn man sich so ein Flugzeugcockpit beispielsweise vorstellt, mehrere Personen in einem Team mit durchaus verteilten Rollen zusammenarbeiten, um in diesem Fall das Flugzeug ans Ziel zu bringen. Und dabei gibt es dann verschiedene Möglichkeiten oder verschiedene Instrumente, die diese Personen im Endeffekt nutzen können, nämlich einmal Anzeigeinstrumente zu Lagebeurteilung und aber auch Steuerungsinstrumente, um darauf einwirken zu können. Und die Anzeigeinstrumente sind bei uns eben sowas wie die Informationszusammenführung über den Ist-Zustand [der Pandemie, J. H.], die letztendlich in einem Dashboard funktioniert. Und die Steuerungsinstrumente sind dann eben die Möglichkeit auszuwählen, […] was wären mögliche [pharmazeutische oder nicht-pharmazeutische, J. H.] Maßnahmen, die wir beschließen könnten, die man einführen könnte. Oder auch, die man wieder auflösen möchte sozusagen, wenn sich die Lage entspannt und dann eben darüber, über die Simulationen einen Eindruck zu gewinnen, wie würde sich das Ganze gegebenenfalls auswirken."

Die Vorteile und Grenzen agentenbasierten Modellierens

Was ist nun aber das Spezifische an ABMS und wie unterscheidet es sich von den ebenso einflussreichen wie strittigen Makromodellierungen? Um Infektionsdynamiken in Szenarien simulieren zu können, liegt Modellen, die auf mathematischen Gleichungen basieren, in der Regel das SIR-Modell (oder eine Variation dieses Modells) zu Grunde. Dieses unterteilt eine Bevölkerung in verschiedene Gruppen: die Empfänglichen (susceptible), die symptomatisch Infektiösen (infectious) und die Genesenen (recovered). Mögliche Infektionsdynamiken können sodann im Hinblick auf die Übergänge zwischen diesen Gruppen modelliert werden. Dem Modell, auf das in der eingangs erwähnten Stellungnahme der Deutschen Gesellschaft für Epidemiologie Bezug genommen wird, liegt beispielsweise das sogenannte SEIR-Modell zu Grunde, in dem zusätzlich die Gruppe der infizierten, aber noch nicht infektiösen Personen (exposed) erfasst wird.

*Epidemiologische Computersimulationen sind Assistenzsysteme, die Entscheidungen nicht selbst treffen, sondern unterstützen; die Entscheidungshoheit haben die Anwender*innen eines Systems.*

In ABMS werden zwar weiterhin diese Gruppen und die Übergänge zwischen ihnen berücksichtigt, allerdings werden die Bewohner einer Stadt oder eines ganzen Landes als individuelle Akteure (= Agenten) modelliert, die entsprechend der ihnen zugewiesenen Handlungslogiken in einer modellierten Umwelt (inter-)agieren. Dieses Modell der Umwelt kann beispielsweise ein digitaler Zwilling[2] einer Stadt sein. Für das Projekt stehen hierfür die ‚Smart-City-Daten' einer Stadt zur Verfügung, die als Kooperationspartner in das Projekt eingebunden ist. Die Handlungslogiken der Agenten werden zumeist entsprechend der als typisch erachteten Tagesabläufe einzelner Subpopulationen modelliert. Dazu wird die zu modellierende Bevölkerung in Alterskohorten, Haushalte oder Berufsgruppen unterteilt, denen jeweils ein bestimmtes Mobilitäts- und Sozialverhalten zugeordnet ist, aus dem sich wiederum Risikobegegnungen ergeben, die zu Infektionen führen können. Für das im Projekt entwickelte

2 Als ‚digitaler Zwilling' wird eine Repräsentation eines Objekts aus der realen Welt in der digitalen Welt bezeichnet.

 https://doi.org/10.14512/tatup.32.1.30

Modell wird in erster Linie das Alter einer Person als relevantes Unterscheidungsmerkmal innerhalb der Bevölkerung erachtet. Das Verhalten der einzelnen Agenten wird entweder in Form von ‚random walks', also als randomisierte Bewegung im virtuellen Raum modelliert oder entsprechend den Methoden verteilter künstlicher Intelligenz (Epstein 2009). In letzterem Fall werden Agenten als rationale Akteure programmiert, d. h. sie verfolgen bestimmte, vorab definierte Strategien und reagieren auf veränderte Umweltbedingungen mit wiederum vorab definierten, veränderten Verhaltensweisen, würden also beispielsweise bestimmte Orte meiden, falls das Ansteckungsrisiko dort zu hoch wäre. Auch ein Verhalten, das unterschiedliche Optionen entsprechend ihrem Nutzen gegeneinander abwägt, kann programmiert werden. Im beforschten Projekt werden sogar abweichendes Verhalten (Maskenverweigerung, illegale Zusammenkünfte etc.) und freiwillige Selbstquarantäne als Handlungsoptionen der intelligenten Agenten vorgesehen. In dieser modellierten Realität wird sodann der Ausbruch einer Infektionskrankheit simuliert, um zu beobachten, wie sich der Erreger infolge der Mobilität und der Interaktionen der Agenten geografisch und demografisch verbreitet und der Status der Akteure zwischen den einzelnen Gruppen wechselt.

Von den vielen Differenzen, die Modelle, welche auf mathematischen Gleichungen basieren, und agentenbasierte Modelle aufweisen, wird zumeist hervorgehoben, dass Agenten untereinander und mit ihrer Umwelt interagieren. Somit kann ihr Verhalten entweder direkt oder vermittelt über die Veränderung von Umweltbedingungen das Verhalten anderer Agenten beeinflussen. Auf mathematischen Gleichungen basierende Modelle können diese Interaktionen und ihre Effekte nicht berücksichtigen, was zu unrealistischen Annahmen und falschen Prognosen zukünftiger Entwicklungen führe. Denn gerade diese wechselseitige Verhaltensanpassung kann zu Systemzuständen führen, die sich vorab nicht vorhersagen lassen. Es entstehen nichtlineare Prozesse und emergente Effekte, die das unbeabsichtigte Ergebnis der Intentionen und Interaktionen der rational handelnden Agenten sind (Weyer und Roos 2017).

Da es sich bei der ABMS folglich nicht um deterministische, sondern um statistische Modelle handelt, bei denen die Beziehungen zwischen den Variablen stochastisch sind, können bei mehreren Durchläufen einer Simulation unterschiedliche Ergebnisse entstehen. Eine solche Varianz lässt sich als Bandbreite von Ergebnissen innerhalb eines Möglichkeitsraums konzipieren. Während sich die Modelle der ABMS also als realistischere Repräsentationen von komplexen sozialen Systemen verstehen lassen, ist das Ergebnis einer durchgeführten Simulation nicht notwendiger Weise mit dem Ergebnis realer sozialer Interaktionen identisch. Dies wird in einem der Interviews als Bruch mit dem Prognoseversprechen gedeutet, das die Erwartungen an Modellierungen und Computersimulationen sowohl in der Öffentlichkeit als auch in der Politik stark prägt (Ioannidis et al. 2022).

> „Weil wir eine Bandbreite haben, spreche ich da nicht so gerne von Prognose, weil wir im Wesentlichen, ja, nicht die Zukunft tatsächlich vorhersagen können. Wir können nur eine Größenordnung letztendlich vorgeben oder die Bandbreite aufmachen zwischen ‚das ist der Worst-Case, den wir beobachtet haben' und ‚das ist der Best-Case, den wir beobachtet haben'. Und mit einer gewissen Plausibilität liegt dann die Realität wahrscheinlich irgendwo dazwischen."

Die Aussage „mit einer gewissen Plausibilität" verweist allerdings bereits darauf, dass unter den Bedingungen von Emergenz und Nichtlinearität diese Bandbreite wiederum nur einen Teil der möglichen Realitäten, d. h. Systemzustände abbildet. Der bereits zuvor zitierte Interviewpartner veranschaulicht dieses Risiko anhand des für stochastische Verfahren paradigmatischen Würfelbeispiels:

> „Wir haben einen Würfel hundert Mal geworfen, wir haben Ergebnisse zwischen 1 und 6 beobachtet, aber wenn wir den Würfel das hunderterste Mal werfen und sozusagen in der Realität den Würfel werfen, wissen wir nicht, welches dieser Ergebnisse kommen wird. Und es kann durchaus sein, dass es ein zehnseitiger Würfel war und wir aber nur 6 von den möglichen Ergebnissen beobachtet haben und das reale Ergebnis im Endeffekt dann sogar außerhalb dessen liegt."

Diese Reflexion der Grenzen der durch Modelle und Simulationen generierbaren Erkenntnisse führt dann schließlich zu einer doppelten Kritik. Zum einen an der öffentlichen Wahrnehmung von Wissenschaft.

> „Aber das wird in der öffentlichen Diskussion, und das kreide ich noch nicht mal den Kolleginnen und Kollegen in der Wissenschaft an, […] natürlich sehr viel kompakter dargestellt und […] bestimmte Grenzen und Einschränkungen fallen dann schnell mal gerne unter den Tisch. Und dann heißt es: ‚Ja, aber die Wissenschaft hat doch prognostiziert, es wird so und so, und die liegen ja alle falsch! Was können die denn eigentlich'?"

Zum anderen wird aber auch das Auftreten und Selbstverständnis von einigen Wissenschaftler*innen problematisiert, was sich durchaus als implizite Kritik an den zu Beginn erwähnten wissenschaftlichen Politikberatungen mit direkten praktischen Konsequenzen verstehen lässt.

> „Und die Gefahr sehe ich an der Stelle dann insbesondere, wenn die wissenschaftliche Kommunikation nicht mehr einfach nur darum geht, was für Erkenntnisse haben wir, sondern wenn es dann auch darum geht, deutliche Beratung mindestens zu machen, oder aber sogar den Diskurs so ein bisschen vorzugeben, indem man sagt ‚unsere Erkenntnisse sagen, wir müssen es so und so machen und das muss jetzt auch so passieren, ansonsten können wir für nichts mehr garantieren'."

https://doi.org/10.14512/tatup.32.1.30

Risiken einer stark vereinfachten Darstellung von Simulationsergebnissen

Welche Rolle spielet nun aber diese proklamierte Abkehr vom Prognoseparadigma der Modellierung und Simulation sowie die daraus abgeleitete Kritik an ungerechtfertigten Formen der Politikberatung für die praktische Tätigkeit im beforschten Projekt? Um diese Frage zu beantworten bietet es sich an, das ‚Dashboard' zu betrachten, das in seiner anvisierten Funktion als Mensch-Maschine-Schnittstelle zugleich Erkenntnisse aus der ABMS darstellen und Möglichkeiten zur experimentellen Erprobung der Effekte von einzelnen Maßnahmen bieten soll. Das Dashboard ist so gestaltet, dass es sowohl den faktischen Ist-Zustand als auch hypothetische ‚Was-Wäre-Wenn-Zustände' in einen sog. pandemic-pressure-score (PPS) ausdrückt, der unter Berücksichtigung zahlreicher Werte und Gewichtungen berechnet wird. Dieser PPS wird für die folgenden vier Dimensionen generiert:

- Gesundheit der Bevölkerung
- Auslastung des Gesundheitssystems
- wirtschaftliche und soziale Belastung
- Belastung kritischer Infrastrukturen

In den Publikationen des Projekts und den von mir geführten Interviews wird nun betont, dass der PPS neben der Inzidenz und der Hospitalisierungsrate auch nicht-medizinische Faktoren erfasst und sich somit in seinem Differenzierungsgrad von ähnlichen Diensten und Tools abhebt. Dies mag zwar für die Darstellung der Ist-Zustände zutreffen. Im Hinblick auf die Frage, wie der nicht-prognostische Charakter und die Grenzen der Simulationsergebnisse reflektiert werden, ergibt sich jedoch ein anderes Bild. Denn der PPS vermittelt gerade nicht die Bandbreite an möglichen Systemzuständen, sondern suggeriert in seiner metrischen Skalierung (der PPS kann einen Wert zwischen 1 und 100 annehmen), dass es sich um ein definitives und exaktes Ergebnis handelt, das letzten Endes prognostischen Charakter hat.

Die problematischen Effekte, die eine solche Wahrnehmung haben kann, lassen sich im Anschluss an die empirische Wirkungsforschung zum Einsatz von Entscheidungsassistenzsystem formulieren (Skitka et al. 1999). Dort wurde festgestellt, dass Anwender*innen den Output der Systeme kaum kritisch hinterfragen und diesen sogar tendenziell als unfehlbar betrachten, was als ‚automation bias' bezeichnet wird. Eben jener Bias führt nun häufig zu zwei Arten von Fehlern: Bei einem ‚Commission-Fehler' folgen Anwender*innen einer fehlerhaften Empfehlung eines Assistenzsystems. Bei einem ‚Omission-Fehler' hingegen, übersehen Anwender*innen kritische Systemzustände, sofern diese von dem System nicht identifiziert werden. Bei der Anwendung von Modellierungen und Simulationen im Pandemiemanagement könnten Commission-Fehler generell dadurch entstehen, dass die verwendeten Daten fehlerhaft sind oder Werte falsch geschätzt wurden. Bei der ABMS könnten zudem falsche Annahmen über das Verhalten von Agenten sowie eine ungenaue Darstellung der Struktur der Bevölkerung und der Umweltbedingungen für das Verhalten der Agenten eine Fehlerquelle sein. Durch die Verfügbarkeit der Smart-City-Daten, anhand derer der Demonstrator entwickelt und getestet wurde, konnte dieses Problem zwar innerhalb des Projekts in den Hintergrund treten. Jedoch sollte gerade angesichts des im Hinblick auf die Verfügbarkeit von Daten sehr voraussetzungsvollen Ansatzes, die Möglichkeit eines fehlerhaften Outputs in das Design des Tools integriert werden, beispielsweise indem die Wahrscheinlich, dass der errechnete PPS falsch ist, immer zusammen mit diesem angezeigt wird.

Das Risiko für Omission-Fehler wird durch die Mehrdimensionalität des PPS zwar verringert, dennoch bleiben kritische Systemzustände, die nicht in einer der vier Dimensionen erfasst werden, für die Anwender*innen unsichtbar. Zudem werden in dem Dashboard nur die Maßnahmen erfasst (und folglich auch durchspielbar), die bisher bereits im Rahmen der Bekämpfung von Infektionskrankheiten eingeführt wurden. Damit begrenzt sich der Horizont des Dashboards jedoch auf den Horizont der Entscheidungsträger, die sich bisher dem Management von Pandemien gewidmet haben. Alternative Ansätze für den Umgang mit Pandemien bleiben qua Design außen vor (Littoz-Monnet 2020).

Mit Blick auf das Design des Dashboards lässt sich noch ein weiteres Problem identifizieren. Die aktuelle Lage und die Auswirkungen von einzelnen Maßnahmen werden nicht nur in einen PPS ausgedrückt, sondern auch anhand eines Ampelsystems visualisiert. Die Ampelfarbe veranschaulicht inwiefern die Gesundheit der Bevölkerung, das Gesundheitssystem, die kritischen Infrastrukturen und die Wirtschaft aktuell belastet sind oder durch die Effekte von Maßnahmen belastet werden würden: Grün steht für eine geringe, gelb für eine mittlere und rot für eine hohe Belastung. Die drei Farben entsprechen jeweils einen bestimmten Bereich des PPS. Dieser Ansatz wurde mit Blick auf die Bedarfe von politischen Entscheidungsträgern in Krisensituationen gewählt. So verweist einer der Interviewpartner darauf, dass dem Projekt von Seiten einer der Partnerstädte des Projekts angetragen wurde, dass deren Bürgermeister „sowas wie ein Ampelsystem" als Entscheidungsgrundlage benötige. Dies ist einerseits durchaus nachvollziehbar, denn angesichts der Fülle an potenziell relevanten Informationen die zur Verfügung stehen, kann eine Entscheidung, insbesondere, wenn sie unter Zeitdruck gefällt werden muss, nur erfolgen, wenn sie nicht alles reflektieren und berücksichtigen muss, was es prinzipiell zu reflektieren und zu berücksichtigen gäbe. Um einen ‚information overflow' zu verhindern, kann ein Ampelsystem also durchaus nützlich und zielführend sein. Andererseits ist diese Komplexitätsreduktion das Einfallstor für implizite Wertungen und Weltbilder der beteiligten Forscher*innen und Programmierer*innen, da die Schwellenwerte, die den Übergang von grün zu gelb und von gelb zu rot bestimmen, sich ja nicht aus dem PPS selbst ableiten lassen, sondern festgelegt werden müssen.

 https://doi.org/10.14512/tatup.32.1.30

Fazit

ABMS ermöglichen im Vergleich zu mathematischen Modellierungen komplexere Darstellungen von Infektionsdynamiken und machen in der Form von ‚Was-Wäre-Wenn-Szenarien' die Auswirkungen von Maßnahmen zur Bekämpfung einer Pandemie antizipierbar. Der emergente und nicht-lineare Charakter von agentenbasierten Simulationen widerspricht jedoch einem Prognoseversprechen, das die öffentliche Wahrnehmung von Modellierungen bestimmt und nicht zuletzt von einem Teil der in der wissenschaftlichen Politikberatung tätigen Modellierer*innen genährt wird. Diese Grenzen der ABMS sollten gegenüber Öffentlichkeit und politischen Entscheidungsträger*innen klar kommuniziert werden und auch im Design von entsprechenden Entscheidungsunterstützungssystemen zum Ausdruck kommen. Dadurch kann überzogenen Erwartungshaltungen entgegengewirkt werden, die folgenschwer sein können. In diesem Sinne gilt auch für ABMS die zentrale Forderung des von Saltelli et al. (2020, S. 484) verfassten Manifests: „Models' assumptions and limitations must be appraised openly and honestly. Process and ethics matter as much as intellectual prowess."

Funding • Funding was provided by the Fritz-Thyssen-Foundation.
Competing interests • The author declares no competing interests.

Literatur

Adam, David (2020): Special report: The simulations driving the world's response to COVID-19. In: Nature 580 (7803), S. 316–318. https://doi.org/10.1038/d41586-020-01003-6

Deutsche Gesellschaft für Epidemiologie (2020): 2. Stellungnahme der Deutschen Gesellschaft für Epidemiologie (DGEpi) zur Verbreitung des neuen Coronavirus (SARS-CoV-2). Online verfügbar unter https://www.awmf.org/fileadmin/user_upload/dateien/covid_19_leitlinien/6.2.pdf, zuletzt geprüft am 03. 02. 2023.

Epstein, Joshua (2009): Modelling to contain pandemics. In: Nature 460 (7256), S. 687. https://doi.org/10.1038/460687a

Epstein, Joshua; Parker, Jon; Cummings, Derek; Hammond, Ross (2008): Coupled contagion dynamics of fear and disease. Mathematical and computational explorations. In: PloS one 3 (12), S. e3955. https://doi.org/10.1371/journal.pone.0003955

Frias-Martinez, Enrique; Williamson, Graham; Frias-Martinez, Vanessa (2011): An agent-based model of epidemic spread using human mobility and social network information. In: 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, S. 57–64. https://doi.org/10.1109/PASSAT/SocialCom.2011.142

Hälterlein, Jens (2020): Die Simulation der Pandemie: Ein Beitrag zur Reihe „Sicherheit in der Krise". Online verfügbar unter https://www.soziopolis.de/die-simulation-der-pandemie.html, zuletzt geprüft am 03. 02. 2023.

Ioannidis, John; Cripps, Sally; Tanner, Martin (2022): Forecasting for COVID-19 has failed. In: International Journal of Forecasting 38 (2), S. 423–438. https://doi.org/10.1016/j.ijforecast.2020.08.004

Littoz-Monnet, Annabelle (2020): Depoliticising through expertise. The politics of modelling in the governance of COVID-19. Online verfügbar unter https://globalchallenges.ch/issue/special_1/depoliticising-through-expertise-the-politics-of-modelling-in-the-governance-of-covid-19, zuletzt geprüft am 03. 02. 2023.

Lorig, Fabian; Johansson, Emil; Davidsson, Paul (2021): Agent-based social simulation of the COVID-19 pandemic. A systematic review. In: Journal of Artificial Societies and Social Simulation 24 (3), 26 S. https://doi.org/10.18564/jasss.4601

Opitz, Sven (2017): Simulating the world. The digital enactment of pandemics as a mode of global self-observation. In: European Journal of Social Theory 20 (3), S. 392–416. https://doi.org/10.1177/1368431016671141

Saltelli, Andrea et al. (2020): Five ways to ensure that models serve society. A manifesto. In: Nature 582 (7813), S. 482–484. https://doi.org/10.1038/d41586-020-01812-9

Skitka, Linda; Mosier, Kathleen; Burdick, Mark (1999): Does automation bias decision-making? In: International Journal of Human-Computer Studies 51 (5), S. 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Weyer, Johannes; Roos, Michael (2017): Agentenbasierte Modellierung und Simulation. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 26 (3), S. 11–16. https://doi.org/10.14512/tatup.26.3.11

DR. JENS HÄLTERLEIN
ist seit 2022 Koordinator des Projekts „Meaningful Human Control – Autonome Waffensysteme zwischen Regulation und Reflexion" an der Universität Paderborn. Er forscht seit mehreren Jahren über die gesellschaftlichen Dimensionen von digitalen Sicherheitstechnologien.

https://doi.org/10.14512/tatup.32.1.30

RESEARCH ARTICLE

Why won't water managers use new scientific computer models?: The co-production of a perceived science-practice gap

Catharina Landström*,[1]

Abstract • The uptake of scientific computer models in water management is challenging. Scientists often face calls to improve stakeholder engagement procedures. However, the involvement of representatives of water management agencies has been common practice in scientific projects for at least a decade. It is therefore questionable whether more stakeholder involvement would lead to greater use of scientific models in water management. This study suggests that computer modeling has historically developed differently in water science and water management. Scientific research has focused on continuous improvement of model process representation, while water management has emphasised usability. Today, the reliance on modeling software packages in water management, exacerbated by the dynamics in the field, mitigates against the adoption of new scientific modeling tools.

***Warum nutzen Wassermanager keine neuen wissenschaftlichen Computermodelle?**: Die Koproduktion einer vermeintlichen Lücke zwischen Wissenschaft und Praxis*

Zusammenfassung • *Die Einführung wissenschaftlicher Computermodelle in der Wasserwirtschaft ist eine Herausforderung. Wissenschaftler sehen sich dabei oft mit der Forderung konfrontiert, die Verfahren zur Einbindung von Interessengruppen zu verbessern. Die Einbeziehung von Vertretern der Wasserwirtschaftsbehörden ist jedoch seit mindestens einem Jahrzehnt bei wissenschaftlichen Projekten gängige Praxis. Es ist daher fraglich, ob eine stärkere Beteiligung von Akteuren zu einer breiteren Nutzung wissenschaftlicher Modelle in der Wasserwirtschaft führen würde. Diese Studie legt nahe, dass sich die Computermodellierung in der Wasserwissenschaft und der Wasserwirtschaft historisch unterschiedlich entwickelt hat. In der wissenschaftlichen Forschung wurde der Schwerpunkt auf eine kontinuierliche Verbesserung der Modelle zur Prozessdarstellung gelegt, während in der Wasserwirtschaft die Benutzerfreundlichkeit im Vordergrund stand. Heute steht die Abhängigkeit von Modellierungssoftware in der Wasserwirtschaft, verstärkt durch die Dynamik in diesem Bereich, der Einführung neuer wissenschaftlicher Modellierungswerkzeuge entgegen.*

Keywords • *computer modeling, water management, stakeholder, science, co-production*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Introduction

Scientists studying water with the use of computer models are today often expected to not only generate new knowledge but also to create new computational tools for use in water management. Research funding agencies ask for scientific projects to achieve societal impact by providing new model-based digital tools for use by water management professionals (Williams 2020). To effectively deliver such tools university scientists have been advised to involve stakeholder representatives from water management organizations directly in research projects (Colosimo and Kim 2016). Despite devising research projects with extensive stakeholder involvement and creating tools that participating professionals find both useful and usable, getting these computer modeling programs adopted in water management has remained a challenge. Recent discussions in scientific journals in water research address this issue and while there is continuing underlining of the importance of improved stakeholder engagement there is also a growing recognition of the importance

* *Corresponding author: catharina.landstrom@chalmers.se*
[1] *Chalmers University of Technology, Gothenburg, SE*

https://doi.org/10.14512/tatup.32.1.36
Received: 26. 08. 2022; revised version accepted: 13. 12. 2022;
published online: 23. 03. 2023 (peer review)

of the institutional context (Wardropper and Brookfield 2022). The present paper contributes to this emerging theme with an exploratory overview of the historical development of distinct computer modeling practices in water science and water management from the 1970s to the present. It is argued that the configuration of actors and their relationships in the relevant governance contexts are key determinants for the adoption of computer modeling tools in water management. In the following the societal co-production of two distinct computer modeling traditions in water science and water management is outlined and the inefficacy of stakeholder participation as a means to transfer computer programs from science to management is clarified.

'Matters of concern' is a social studies of science concept that was introduced by Latour (2004) to distinguish between scientific propositions accepted as matters of fact and contested scientific knowledge claims. Making the distinction in this way opened for new lines of inquiry and of particular interest for the present study is the practical and temporal dimensions. Matters of concern directs attention to the socio-material processes through which some elements are stabilised and taken for granted over time. The emphasis is on actions, on the practices in which some perspectives become self-evident to the people involved. Applying this concept to the evolution of computer modeling of water focuses attention to the actors' discussions about what they want their models to do. In the next section of the paper this concept informs a historical outline of the development of computer models in water science and water management.

In the present paper the reference point is water related decision-making more broadly.

Aiming to capture processes evolving through time qualitative document analysis (Bowen 2009) is the primary method used to develop the argument of this paper. Relevant texts have been traced and accessed in online publication databases. Such an effort can only ever be partial, and the findings presented here are not intended to be exhaustive, the ambition is to suggest a critical social studies of science perspective that may inform further investigation of this complex issue.

The paper begins with an outline of the key concepts which are subsequently used to organise the argument in three sections. The first substantive section details the historical evolution of computer modeling in water science and water management drawing on the notion 'matters of concern' (Latour 2004) to emphasise practice. The second empirical section examines the way water modeling is shaped in a water governance space (Lange and Cook 2015), with the United Kingdom (UK) as an example. Finally, the divergence of scientific computer modeling and computer modeling in water management is discussed as the outcome of a wider process of co-production (Jasanoff 2004).

A brief note on key concepts

The objects that this paper pivots on are computer programs that make it possible for users to analyze the aspects of water systems that are of interest to them by simulations. Such programs are referred to with many different terms, depending on who is talking. For example, scientists often talk about them as models and about modeling as a process beginning with theory, concepts and mathematical formulas, that they express in computer code. In contrast water management professionals may speak of models with reference to off-the-shelf software packages, that can include several computer programs, which they apply to a water problem in a particular location. This paper does not prescribe a particular use of the terms but follows the actors and tries to clarify what is meant in each reported conversation.

Computer modeling of water plays a role in a wide range of societal practices, here the focus is on the scientific study of water as resource and environmental feature and on the management of it, including flooding, drought, and water quality. Scientific research on water is distributed among many scientific disciplines of which hydrology is a major field but other fields, such as geology and chemistry, also model water scientifically. Water management is also distributed across many actors in society, ranging from national government departments to local authorities and private businesses, which operate on varying geographical scales and with distinctive legal responsibilities. To capture this complex multitude, we adapt the notion of drought governance space used by Lange and Cook (2015). Lange and Cook aimed to capture the complex relationships between the actors involved with drought decision-making in the UK but in the present paper the reference point is water related decision-making more broadly.

The modified notion of 'water governance space' indicates networks of heterogenous actors who are linked together by issues involving water, such as flooding. The water governance space encompasses both science and management and the relationship between them and with other institutions and organizations.

'Co-production' is a well-known and widely used concept in science studies and beyond. Sheila Jasanoff's (2004) elaboration of the notion emphasises historical process and institutional dynamics. This paper embraces Jasanoff's idea that science and society co-produce knowledge and expertise that decide the ways in which environmental challenges, such as drought and flooding, are addressed at a particular time in a specific place. The final section of this paper deploys the notion of co-production in a critical discussion of the discursive emphasis on direct in-

volvement of individual stakeholder representatives in scientific research as the most important measure to increase the uptake of scientific models and modeling tools in water management.

Distinct matters of concern

Computer modeling was adopted as a technique for investigating water in the early days of computing. In a retrospective article Keith Beven, a still active professor emeritus in hydrology, recalls that his first model was "programmed in Algol and physically existed as a pack of punched cards that needed to be fed into a card reader every time a run was made" (Beven 2019, p. 1481). Drawing on his long experience he insists that the reason for water scientists to model has always been, and remains, to "test the understanding of how a hydrological system might function" (Beven 2019, p. 1486). The main purpose of Beven's article is to discuss the priorities for hydrological modeling, the retrospective demonstrates the continuity of the concern with model improvement through the decades. Beven's construction of a genealogy underpins his identification of important issues in hydrological modeling – deciding if a model is fit for purpose, improving process representation and to create "models of everywhere" – as perpetual preoccupations in water science. For the present purpose it does not matter whether other water scientists in academia agree that these are the most important issues, what is of interest is that we can discern distinctly different matters of concern in water management.

The historical evolution of computer modeling in water management can be traced via instruction manuals for water engineers and other texts with a practical purpose, which were written as it became possible to run computer programs on office desk top computers. Potential model users needed information about how to deploy the computer programs. In a 1982 guidance text the US Office of Technology Assessment (OTA) explained that:

> "While many of the economic and social factors in water resource decisions cannot be fully enumerated, models can be used to integrate the available data, and provide estimates of future effects and activities. Such estimates are highly useful in evaluating the consequences of different resource policy options, and are often less expensive than conducting comprehensive surveys and using other traditional approaches" (OTA 1982, p. 6).

This quote explains the purpose of computer modeling in water management in a concise way – to bring the available quantitative information together and enable estimation of potential consequences of actions. It also brings practical considerations to the forefront by noting that computer modeling is often cheaper than other ways of generating the knowledge needed. Further, the OTA also identified challenges to the successful use of computer models in water management:

> "Presently, model development has outstripped corresponding support for models. In the past, model developers have put a premium on developing models, while support for models – documentation, validation, dissemination, user assistance, and maintenance – has been neglected. Often, resources are focused on development, but are unavailable for support activities" (OTA 1982, p. 9).

This shows that maintenance of computer programs and user support emerged as matters of concern in the early days of water management modeling. This was not an issue for university scientists who wrote their own computer programs to address research question they had formulated in relation to scientific discourses. In contrast water management required computer programs that could be used by professionals with different disciplinary backgrounds, whose work included ensuring that good quality water was delivered to households and businesses, that properties were protected from flooding and that there would be sufficient supply of water in times of drought.

Both science and water management in the 1970s and 80s called for improvement of computer models, but with different matters of concern. Scientists sought to achieve more accurate representations of natural processes which emphasized the theoretical and mathematical models encoded in computer programs. In contrast, the concern of water management was to address real-world problems, thus prioritising reliability and usability of computer programs for different users. While reliability requires comparison of different aspects of computer programs, including the scientific quality, usability points in the direction of standardisation of computer programs in software. In the 1990s, a guide for engineers remarked that:

> "The model user community has grown dramatically, particularly in regard to local public agencies, private consulting firms and other non-federal users. Most of the water management models cited throughout this report include user-friendly executable (ready-to-run) versions for desktop computers. Essentially everyone in the water management community now has convenient access to the computer hardware needed to run the available software" (Wurbs 1994, p. 4).

The quote is from a US publication, but computer modeling had become a key tool in European water management too by the 1990s (Seibert and Bergström 2022). Worldwide the issue of user-friendly software featured prominently in discussions of water management, and this appears to have impacted the development of computer modeling programs. The same US guide for water engineers notes that:

> "Model development in recent years has been characterized by an emphasis on interactive user interfaces oriented toward using advances in computer technology to make models more convenient to use. […] Enhanced user interfaces

 https://doi.org/10.14512/tatup.32.1.36

> have been a key consideration incorporated in the development of newer water management models and have been recently added to a number of older models" (Wurbs 1994, p. 13).

The emphasis on the need to improve user interfaces and user support in guidance texts from both 1982 and 1994 shows that usability is a long-running matter of concern in computer modeling for water management. Scientific computer modeling did not evolve in a way that could satisfy this need, instead there has been a proliferation of models implemented in computer codes which continues in the present. Some scientists view this as a problem.

> "Today the plethora of available models has grown beyond any possible limit and the need for accommodating under a unifying view and reconciling the different approaches has become of great priority" (Todini 2011, p. 73).

Others see good reasons for creating new models:

> "One of the key drivers for the pronounced model diversity in hydrology is undoubtedly the wide range of model applications […] that all require appropriate modeling […]. Two well-accepted characteristics that models should exhibit are parsimony and adequacy to the problem at hand, that is, a model should not be more complex than necessary and should be fit-for-purpose […]" (Horton et al. 2021).

Parsimony requires that a computer model does not include redundant algorithms that may slow down its running and affect stability, adequacy is about the faithful representation of the specific local process. These are scientific criteria that drive ongoing development of new models, other considerations are:

> "Researchers […] are, in fact, very keen on using local models, either developed in Switzerland […] or even at their own research institute. […] This model-institute link is likely to be one of the main causes of the existence of so many hydrological models, since each research group develops its own tools" (Horton et al. 2021).

Water scientists test their explanations of water system behaviour by creating computer programs and running models. That scientists want to use their own computer program codes, or codes that their research team use speaks to the way in which they develop trust in a model. Familiarity confers transparency, scientists see through trusted computer programs to the physical processes they investigate.

When water scientists have completed research projects they often move on to a new project, addressing different questions. In the new project they may create a new computer model or radically rewrite a computer program they have used previously, already in the late 1990s it was noticed that:

> "The majority of the thousands of water-related computer models reported in the literature […] were developed to support a particular study, which was eventually completed and the model shelved" (Wurbs 1998, p. 192).

Computer programs written to address the questions driving particular research projects are rarely used again in the original format, hence the code is not tested for reliability across contexts or made usable for non-scientists. Scientific practice does not push the development of computer programs towards resolving matters of concern in water management, even when they achieve better process representation.

Computer modeling in the water governance space

Water management involves more than knowledge about the physical behaviour of water, a 1990 guide for engineers and other practitioners explains that "the planner must effectively consider the political aspects of the project, and he [sic!] must be clear about needs, goals, objectives and expectations resulting from the socio-economic and cultural system and being influenced by the infrastructure" (Dyck 1990, p. 4). That many elements influencing water decision-making cannot be captured in computer models has been recognised for decades.

> "[… M]athematical programming techniques cannot provide a unique optimal solution for a water resources system. For long-term planning and management, methods are required which reflect the complex, interactive, and subjective character of the decision-making process, taking into account the experiences of the decision-makers […]" (Dyck 1990, p. 5).

In real-life water-related decision-making other people than modelers set the agenda and computer modeling is only one element of many. This raises the question of how to incorporate modeling in decision-making. One approach is to consider computer models as a type of decision-support tools.

> "The application of decision methods and the integration of such methods with hydrological modelling systems for use in water resources control and management will have also to be advanced. At this stage the hydrological model becomes integrated in the new kinds of architectures and paradigms (of object orientation and agent orientation) that are now becoming established in other fields" (Refsgaard and Abbot 1996, p. 14).

The notion of decision-support spreading through the water science literature in the 1990s remains central today as the demand on scientific projects to show impact through uptake of scientific modeling in water management is often met by commit-

https://doi.org/10.14512/tatup.32.1.36

ments to contribute model-based decision-support tools. A 2022 special issue of the Journal of Hydrology focuses on the challenges of creating decision-support tools that are adopted in water management. Interestingly, most papers in the special issue discuss improvements of process representation in new decision-support tools (Wardropper and Brookfield 2022).

In the decade around the turn of century 2000 software for water management modeling had multiplied and continued to do so. Creating modeling software for use in water management has become a specialized technical field, done by multidisciplinary expert teams in businesses operating on global markets. These expert businesses provided modeling software packages as off-the-shelf products for purchase, offering user support and regular upgrades. Written advice was needed for what to consider when investing in a software package:

> "Generalized models should be convenient to obtain, understand, and use. They should also work correctly, completely, and efficiently. Documentation, user support, and user-friendliness of the software are key factors in selecting a model for application. The extent to which a model has been tested and previously applied in actual studies is also an important consideration" (Wurbs 1998, pp. 190–191).

In the time that has passed since advice was directed to individual professionals selecting a software package to use when modeling for water management the number of such packages and the cost of using them has multiplied. It is no longer a decision that can be taken by an individual in an organisation that has financial and legal obligations. Today the complexity of water governance requires standardization of modeling software use, UK flood risk management provides an example of how this can be done.

All UK key actors in the water governance space use computer models – the water utility companies have modeling experts who use software packages to generate actionable knowledge to manage water resources, a task including modeling water supply, sewage, run-off, drought, flooding, water quality and more. The UK Environment Agency (EA) – the regulatory authority for water as a natural resource that needs to be used in a sustainable manner – models all processes affecting rivers and lakes, including biodiversity. The EA operates under the auspices of Defra (Department for Food, Environment and Rural Affairs) the government agency responsible for water management nationally, which draws on modeling to develop policies for achieving national objectives such as a certain level of flood protection for all properties. Also important in relation to modeling are the expert consultancies who carry out many modeling projects for private businesses, government agencies and local authorities. For UK water management to work – delivering water to households and businesses, removing foul water (including sewage and road runoff), mitigating flood risk, ensuring sufficient river flows for aquatic and riverine ecology to survive and flourish, and so on – everybody must act in concert. In this context computer programs must be stable and transparent.

The need for stability and transparency of computer models in water management is, in the UK, satisfied through benchmarking of available software packages. Since the 1990s the EA, under the auspices of Defra, has undertaken benchmarking projects in collaboration with major consultancy firms and university scientists, to compare the market leading water modeling software packages. The importance of the benchmarking is visible in the EA guidelines for modeling in flood risk management (often carried out by consultancies commissioned by local authorities).

> "You must be able to demonstrate that the software you choose is suitable for the intended use. […] If the tests are done independently, the EA will need to review the results before you use the software for a project" (EA2021).

While everybody is free to use any model that they would like, comparing a new computer modeling program with the benchmarked software would require extensive effort. The water management modeler who wanted to use a non-standard modeling program would have to prove that it performs at least as well as the approved packages on the list provided by the EA:

- "To produce a 1D model, you can use: Flood Modeller, ESTRY, HEC-RAS, InfoWorks ICM, MIKE FLOOD
- To produce a 2D model, you can use: Flood Modeller, TUFLOW, HEC-RES, InfoWorks ICM, MIKE FLOOD, JFlow®
- To produce a 1D2D model, you can use: Flood Modeller, TUFLOW Classic, HEC-RES, InfoWorks, MIKE FLOOD, JFlow®" (EA 2021)

All but two of the approved software packages are proprietary. Developed by the US Corps of Engineers HEC-RAS and HEC-RES are free to download and use since they were developed with federal funding. The other approved packages were developed by expert consultants, often starting with a computer program written by university scientists as a PhD or post doc project.

Modelers in science and in water management talk about water in much the same way.

https://doi.org/10.14512/tatup.32.1.36

The UK example shows how models can be stabilized and controlled as software packages in the water governance space. In other countries the required stabilization of modeling software may be achieved in other ways. One example is through close long-term collaborations between creators and users of modeling software in Sweden where SMHI (the Swedish Meteorological and Hydrological Institute) has provided modeling services and software for water management to municipal and regional authorities since the 1970s. In the Netherlands the research institute Deltares develops and supplies an array of software tools for use in water management. Regardless of how modeling software is stabilized in a particular water governance space, it is a process that runs counter to the continuous creation of new computer models in academic science.

Concluding discussion: a co-produced gap

Recently science funding agencies have voiced concerns about the creation of computer models in water science that are only used in the funded project. This is considered a problem and scientists have responded to this by promising to deliver decision-support tools for use in water management through increased stakeholder engagement. However, the historical outline of modeling in water science and water management in the previous suggests that increasing, the already common stakeholder involvement, will not lead to more uptake of new scientific computer models in water management.

The document study showed that although computer models are central in both water science and water management the matters of concern arising in these practices that drive model development differed from the outset. Scientists work in contexts where they are free to develop new scientific computer modeling programs that enable them to answer their research questions. A successful scientific computer model points the scientists to the aspects of the studied process in nature that are still poorly understood. In contrast, water management practices require computer programs that can be used to address the same questions in different physical contexts by professionals with diverse expertise. A successful water management model is easy to use, reliable across contexts of use and it produces information that is appropriate for the decision-making situation. The differences between what is a useful computer model in science and in water management explains why water scientists cannot deliver models or modeling tools directly to users in water management. The requirements on water management models have prompted development of modeling software packages. A successful scientific model needs further development to become usable software.

The differences between what computer programs need to do in the two domains can be understood as resulting from historical and institutional co-production (Jasanoff 2004). Water science and water management have evolved as distinct practices conducted in different societal institutions. Science provides a framework for experimentation and invention within the water governance space while water management is responsible for controlling water as a common good.

Interestingly, modelers in science and in water management talk about water in much the same way which could contribute to the positive valuations of stakeholder involvement in research projects by both scientists and stakeholder representatives. This positive interaction does not often lead to the desired transfer of new models into water management practice. An individual stakeholder representative's interest in a new computer modeling program does not make it possible for her or him to start using it in their everyday work that is defined by the requirements of institutional decision-making. Water management computer modeling software must be transparent and trusted by the institutions and organizations in the water governance space, who are affected in by the decisions. The UK example shows how the selection of software packages for modeling in flood risk management can be controlled by institutional actors. Although the constraints of model choice for decision-making may be less explicit in other places and in relation to other issues the time-consuming comparison of modeling software packages and the need to provide user support, technical upgrades and so on mitigate against the adoption of novel scientific computer models in water management.

Funding • This work received no external funding.
Competing interests • The author declares no competing interests.

References

Beven, Keith (2019): How to make advances in hydrological modelling. In: Hydrology Research 50 (6), pp. 1481–1494. https://doi.org/10.2166/nh.2019.134

Bowen, Glenn (2009): Document analysis as a qualitative research method. In: Qualitative Research Journal 9 (2), pp. 27–40. https://doi.org/10.3316/QRJ0902027

Colosimo, Mark; Kim, Hynnok (2016): Incorporating innovative water management science and technology into water management policy. In: Energy, Ecology and Environment 1 (1), pp. 45–53. https://doi.org/10.1007/s40974-016-0013-z

Dyck, Siegfried (ed.) (1990): Integrated planning and management of water resources. (Guidance material for courses for engineers, planners and decision-makers). Available online at https://unesdoc.unesco.org/ark:/48223/pf0000155111, last accessed on 04. 02. 2023.

EA – Environment Agency (2021): Hydraulic modelling. Best practice (model approach). Available online at https://www.gov.uk/government/publications/river-modelling-technical-standards-and-assessment/hydraulic-modelling-best-practice-model-approach, last accessed on 04. 02. 2023.

Horton, Pascal; Schaefli, Bettina; Kauzlaric, Martina (2021): Why do we have so many different hydrological models? A review based on the case of Switzerland. In: WIREs Water 9 (1), pp. 1–32. https://doi.org/10.1002/wat2.1574

Jasanoff, Sheila (ed.) (2004): States of knowledge. The co-production of science and social order. Abingdon: Taylor & Francis.

Lange, Bettina; Cook, Christina (2015): Mapping a developing governance space. Managing drought in the UK. In: Current Legal Problems 68 (1), pp. 229–266. https://doi.org/10.1093/clp/cuv014

Latour, Bruno (2004): Why has critique run out of steam? From matters of fact to matters of concern. In: Critical inquiry 30 (2), pp. 225–248. https://doi.org/10.1086/421123

OTA – US Office of Technology Assessment (1982): Use of models for water resources management, planning and policy. Available online at https://ota.fas.org/reports/8233.pdf, last accessed on 04. 02. 2023.

Refsgaard, Jens; Abbot, Michael (1996): The role of distributed hydrological modelling in water resources management. In Michael Abbot and Jens Refsgaard (eds.): Distributed hydrological modelling. Dordrecht: Kluwer Academic Publishers, pp. 1–16. https://doi.org/10.1007/978-94-009-0257-2_1

Seibert, Jan; Bergström, Sten (2022): A retrospective on hydrological catchment modelling based on half a century with the HBV model. In: Hydrology and Earth System Sciences 26 (5), pp. 1371–1388. https://doi.org/10.5194/hess-26-1371-2022

Todini, Ezio (2011): History and perspectives of hydrological catchment modelling. In: Hydrology Research 42 (2–3), pp. 73–85. https://doi.org/10.2166/nh.2011.096

Wardropper, Chloe; Brookfield, Andrea (2022): Decision-support systems for water management. In: Journal of Hydrology 610, p. 127928. https://doi.org/10.1016/j.jhydrol.2022.127928

Williams, Kate (2020): Playing the fields: Theorizing research impact and its assessment. In: Research Evaluation 29 (2), pp. 191–202. https://doi.org/10.1093/reseval/rvaa001

Wurbs, Ralph (1998): Dissemination of generalized water resources models in the United States. In: Water International 23 (3), pp. 190–198. https://doi.org/10.1080/02508069808686767

Wurbs, Ralph (1994): Computer models for water resources planning and management. Available online at https://usace.contentdm.oclc.org/digital/collection/p16021coll2/id/3719/, last accessed on 04. 02. 2023.

DR. CATHARINA LANDSTRÖM
holds a position as Associate Professor in STS at Chalmers University of Technology since 2018. Before that she worked in transdisciplinary water research projects in the UK and came to specialize in the social study of environmental computer modeling positioned in between science and society.

https://doi.org/10.14512/tatup.32.1.36

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom

RESEARCH ARTICLE

Combining behavioral insights with artificial intelligence: New perspectives for technology assessment

Lilla Horvath[1], Erich Renz[1], Christian Rohwer*,[1], Daniel Schury[1]

Abstract • Policy decisions concerning technology applications can have far-reaching societal consequences. Rationality-enhancing procedures are thus essential to ensure that such decisions are in the best interest of society. We propose a novel framework addressing this challenge. It combines a structured approach to decision-making, the mediating assessments protocol (MAP), with artificial intelligence (AI) methods to mitigate human bias and handle uncertainty in a normative manner. We introduce the steps for implementing MAP and discuss how it can be complemented and improved by AI methods such as dynamic programming, reinforcement learning and natural language processing. As a potential practical application, we consider the construction of a new wind park in a community and highlight critical aspects warranting special caution.

***Über die Verbindung von Erkenntnissen der Verhaltensforschung mit Methoden künstlicher Intelligenz**: Neue Perspektiven für die Technikfolgenabschätzung*

Zusammenfassung • *Politische Entscheidungen in Bezug auf Technikanwendungen können weitreichende gesellschaftliche Folgen haben. Rationalitätsfördernde Verfahren sind daher unerlässlich, um sicherzustellen, dass die Entscheidungen im Interesse der Gesellschaft getroffen werden. Wir stellen hier eine neue Methode für ein solches Verfahren vor. Unser Ansatz kombiniert ein strukturiertes Verfahren zur Entscheidungsfindung, das sogenannte Mediating Assessments Protocol (MAP), mit Methoden der künstlichen Intelligenz (KI), um den Einfluss menschlicher Voreingenommenheit zu reduzieren und Unsicherheiten normativ zu handhaben. Wir beschreiben die Implementierung von MAP und erörtern, wie dieses von KI-Methoden wie der dynamischen Programmierung, verstärkendem Lernen und der automatischen Verarbeitung natürlicher Sprache profitiert. Anhand eines Beispiels zur Errichtung eines Windparks in einer Kommune veranschaulichen wir unseren Ansatz und zeigen kritische Aspekte auf, bei denen besondere Vorsicht geboten ist.*

Keywords • *artificial intelligence, behavioral economics, human bias, policy decisions, uncertainty*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Introduction

The future of a state and its citizens can be impacted significantly by the introduction of new technologies as well as the termination or change of existing technologies. Therefore, the associated political decision-making processes are crucial: To benefit society, policy measures must be informed by a thorough assessment of the possible consequences of a technology application. We address two key factors that, in our view, complicate this undertaking. First, the consequences of a technology application are, in general, of a probabilistic nature: Various outcomes could occur with different probabilities. These probabilities and the outcomes are often subject to imprecision, either because they are inherently only partially accessible or because relevant data are missing. Therefore, most policy decisions are imbued with uncertainty. Second, while technology assessment can be carried out by independent experts, policy measures are implemented by political decision-makers who might be bound by the agendas of their parties, constrained by their own cognitive biases (e.g., herd mentality, which means that people tend to copy the behavior of those with whom they feel connected, even

* *Corresponding author: christian.rohwer@pd-g.de*
[1] *PD – Berater der öffentlichen Hand GmbH, Berlin, DE*

https://doi.org/10.14512/tatup.32.1.43
Received: 26. 08. 2022; revised version accepted: 13. 01. 2023;
published online: 23. 03. 2023 (peer review)

though they would act differently if they were to decide on their own) or limited through time and available resources.

We propose a framework addressing both factors *uncertainty* and *human bias and constraints* in order to facilitate better policy decisions.

Our framework combines structured decision-making protocols (Kahneman et al. 2021) with quantitative methods from the field of artificial intelligence (AI) (Russell and Norvig 2021). The decision-making protocol employs the Delphi method (Beiderbeck et al. 2021): Political decision-makers are provided with reports as a basis for all policy-related discussions. This step is followed by discussions of individual assessments that feed into the final decision, and a consultative process culminating in consensus-based, independent and transparent policy decisions. While this decision-making protocol helps to minimize human bias and constraints, in order to improve its result from a normative standpoint, reports provided to political decision-makers should include action plans that account for uncertainty in a systematic manner. To this end, we propose that qualified experts employ AI methods such as reinforcement learning, dynamic programming, Bayesian modeling and natural language processing in order to enhance the quality of reports being provided to decision-makers. These tools offer a formal basis for handling uncertainty from a normative perspective and promote the processing of growing amounts of data by pre-filtering information.

The Office of Technology Assessment at the German Bundestag (TAB) is responsible for technology assessment in Germany at the federal level. Its tasks include analyzing the impact of scientific and technological developments as well as the associated opportunities and risks from social, economic and ecological standpoints. Based on these analyses, committees and members of parliament receive recommendations for actions by TAB. However, on the municipal level, city councils cannot rely on analyses by TAB. Furthermore, on the municipal level, action plans regarding new technologies have to accommodate specific local conditions. With our framework, we address municipal political decision-making.

Framework part 1: mediating assessments protocol (MAP)

In this section, we make the following assumption: Technology assessment in the public sector spans from policy recommendations to specific decisions which either

a) introduce a new technology application,
b) terminate an existing technology application, or
c) change an existing technology application.

All three of these prototypical decisions have societal implications concerning opportunities and risks. We propose the introduction of the mediating assessments protocol (MAP) (Kahneman et al. 2019, 2021) for technology assessments. MAP is a structured approach to strategic decisions developed by Kahneman et al. (2019). They describe strategic decisions as 'evaluative judgments' in which decision-makers break down multi-layered information to choose among options based on rankings or to embark on a new initiative based on a binary yes-no decision.

Our framework accounts for uncertainty and human bias to improve policy decisions.

We argue that the MAP-methodology for technology assessments should be supplemented by methods from the field of AI in order to formally deal with uncertainty in the above-mentioned decisions. The purpose of MAP is to reduce human decision errors such as those resulting from cognitive biases (Kahneman 2011), from noise due to a variation in judgments that should be similar, or from noise due to attention to seemingly irrelevant factors.

Political decision-making at the municipal level differs from that at the state or federal level. One reason for this is that at the municipal level, the ruling majority is often heterogeneous because different local interests are represented directly and central party positions tend to be of lower importance. Therefore, to make majority decisions, different political interests in the municipal council have to be aligned. However, reaching consensus can be complicated because people tend to misinterpret data-based facts (Stolwijk and Vis 2020) or bias these towards their political beliefs (Alesina et al. 2020). To overcome these pitfalls, we propose the MAP-framework as detailed below.

In the kick-off meeting, the decision-making body (e.g., the municipal council) defines specific evaluation dimensions of the technology to be assessed. For example, if the decision is related to constructing community wind parks in order to increase the share of local green energy, evaluation dimensions such as social acceptance, switching and acquisition costs for the community or overall impact on sustainable community goals could be included. Next, experts (either internal employees or external consultants) prepare an objective and independent report on each evaluation dimension, also using AI methods (see next section). For each evaluation dimension experts should aim to answer the question 'Do the findings in the evaluation dimension (e.g., social acceptance, switching costs, overall impact on sustainable community goals etc.) support or oppose the construction of wind parks?'

https://doi.org/10.14512/tatup.32.1.43

For each evaluation dimension, it is important to work out a 'base rate'. Returning to our wind park construction example, for the evaluation dimension 'likelihood of achieving local communal sustainability goals' the base rate is given by the percentage of wind energy in those communities that have already reached similar sustainability goals. In addition, a 'reference class' has to be determined for each evaluation dimension. In our example, this refers to a group of comparable communities in terms of, e.g., size and demographics. Both base rate and reference class for a given dimension are used to generate 'relative judgements', e.g., 'within 100 comparable communities our community ranks no. 30 based on how close it is to reaching the target wind energy percentage, i.e., the base rate'.

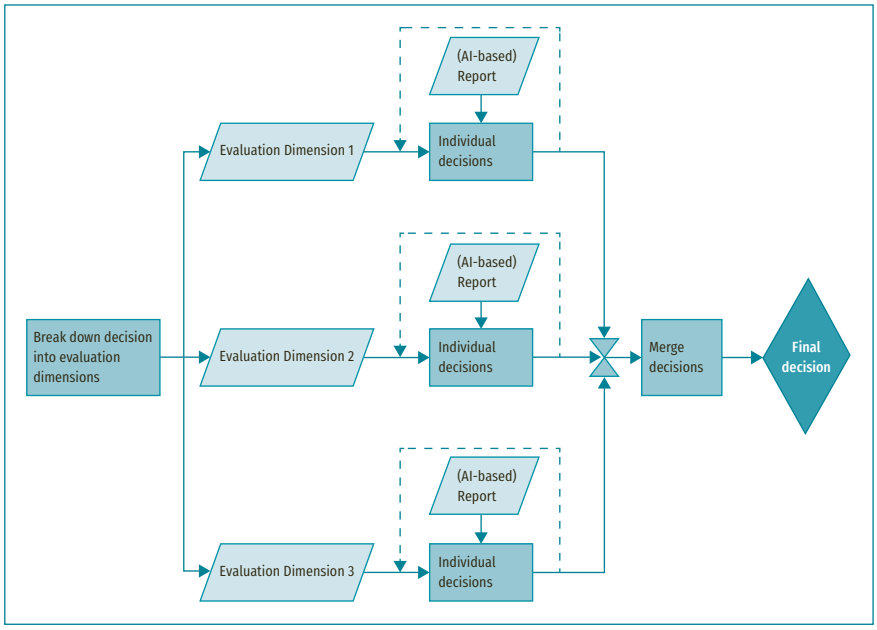


Fig. 1: Schematic flow chart of MAP. Rectangles represent actions; diamonds represent input. The dashed lines illustrate an (optional) repetition of the individual decisions phase. *Source: authors' own compilation*

Experts for an evaluation dimension should assess their dimension independently to minimize the risk of being influenced by other experts. In the event of staff shortages, individual employees could be assigned multiple assessment tasks. In this case, the evaluation dimensions must be clearly delineated so that the quality and objectivity of the analysis does not suffer from possible influence by a previous evaluation dimension that has similar characteristics. When experts report, it is important to include statements on information deficits, but also noteworthy risks for a possible failure of the project, so that they can be taken into account by the decision-making body in a final assessment. Upon completion of the experts' reports for each evaluation dimension, these are forwarded to the decision-making body ahead of the scheduled meeting.

In the decision-making meeting, the decision-making body is likely to be confronted with both positive and negative evaluation outcomes. The body should consider each evaluation dimension independently as a separate discussion item. At this point – or at the very beginning, when evaluation dimensions are defined – the body should agree on a weighting of individual dimensions (e.g., acquisition costs have more weight in the overall evaluation than another dimension).

On the day of the decision meeting, experts summarize key points of each evaluation dimension. Then, each member from the decision-making body votes individually per dimension. The evaluation outcome is used as a guideline, which the member of the decision-making body can agree with or deviate from. Voting takes place anonymously to secure independent individual decisions. While there may be quick agreement on some points, other issues are discussed more vigorously and different positions are put forward. The decision-making body votes again at the end of the debate on an evaluation dimension. It can be assumed that the level of agreement in a second round of voting will be greater than in the first round of voting. This procedure is repeated for each evaluation dimension until a final decision is reached. All mean values for the individual evaluation dimensions are presented. Percentile scales provide a suitable basis for voting (e.g., 'In your opinion, how likely is it that the target wind energy percentage is reached within 1 year on a scale from 0 to 100 %? Within 5 years?'). Based on transparent and data-based evaluations, the committee finally discusses the technology case and votes on how to deal with it. Figure 1 summarizes the flow of MAP.

Framework part 2: AI methods

In this section, we consider methods from the field of AI that could help to guide policy measures pertaining to technological change in the face of uncertainty. These methods can be added to the experts' toolbox for creating reports and should be implemented by AI practitioners. Specifically, we outline two classes of algorithms that have been employed to tackle problems imbued with uncertainty that require step-by-step decisions: dynamic programming and reinforcement learning. To highlight that dynamic programming was developed within the field of operations research (Bellman 2010) we avoid here the term model-based reinforcement learning (Sutton and Barto 2018) which often is used in AI research to refer to dynamic programming. We conclude this section by discussing how natural language processing algorithms could offer further support for this undertaking by extracting relevant information from large text-based datasets.

Both dynamic programming and reinforcement learning seek to identify the action that promises the highest cumulative reward in the long run for each possible situation that might be encountered for a given problem. These algorithms thus offer normative tools for inferring optimal courses of action in sequential decision-making problems under uncertainty and therefore provide potentially valuable tools for enhanced decisions regarding the introduction, termination or change of technology applications. Given that these algorithms optimize action sequences, they can also be used to guide the step-by-step implementation of policy measures. We provide here a brief overview of dynamic programming and reinforcement learning; for detailed discussions of the topics we encourage interested readers to consult Bertsekas and Tsitsiklis 1996; Sutton and Barto 2018; Wiering and Otterlo 2012.

To find an optimal action sequence, both dynamic programming and reinforcement learning algorithms rely on a recursive definition: The best action in a given situation – formally denoted by 'state' – is the action for which the sum of the immediate reward perspective and the maximum longterm reward perspective as captured by the optimal value of the expected next state is maximal. The optimal value of the expected next state is given by the maximum overall reward perspective from that expected state. Dynamic programming algorithms put this recursive definition to use by computing the best action. This, however, requires that the decision-maker – formally referred to as 'agent' – has full knowledge about the probabilistic dynamics of the problem environment, i.e., a probabilistic representation of the consequences of a given action in a particular state for rewards and new states. Standard dynamic programming algorithms employ this knowledge to work their way back from terminal to initial states and can thus deliver optimal solutions beforehand. In contrast, reinforcement learning algorithms require no knowledge about the probabilistic dynamics of the problem environment; instead, the best action is identified by repeatedly interacting with a (simulated) instance of the problem environment and thereby gathering experience with the reward perspective of state-action pairs.

Both dynamic programming and reinforcement learning algorithms have many variants; a particularly important class of these complements the standard schemes with Bayesian methods. Such approaches are indeed essential if the optimal solution is sought for a problem environment where certain components such as the state or the probabilities governing the state transitions and reward emissions are only partially observable. In such problem environments the best course of action must strike an optimal balance between exploration, i.e., expanding the knowledge about the partially observable components of the problem environment, and exploitation, i.e., reward gathering by harnessing the accumulated knowledge. It goes beyond the scope of the present work to discuss the utilization of Bayesian methods in detail (for a standard reference on these methods see Bertsekas 2019; Wiering and Otterlo 2012). Therefore, below, we will give an example for a problem structure in which it is assumed that all components can fully be observed.

Dynamic programming and reinforcement learning have found a host of applications including finance, robotics, gaming and autonomous driving. However, to the best of our knowledge, they have not been used in aiding technology assessment. Incorporating these methods to facilitate better policy decisions for technology applications would require that the building blocks of relevant problems can readily be mapped onto the terminology of states, actions, rewards, state transitions and reward emissions. As in most application areas, this undertaking necessitates substantial domain knowledge and manual fine-tuning. To illustrate this, let us return to our example of the wind park construction, where a relevant problem is to find measures that seek to positively influence social acceptance. In this problem, a first step is to consider main concerns regarding the construction of wind parks, such as the visual impact on the landscape, noise or the impact on the local ecosystem (Leiren et al. 2020). Choosing to prioritize a particular concern can be viewed as a possible action following from the initial state. Each such action yields a certain reward, which in this problem corresponds to a public reaction, and leads to a new state where a new set of actions becomes available. Figure 2 shows a schematic of the problem structure with hypothetical reward and state transition dynamics for this particular example. To make this concrete, a rigorous mapping of states and parameters could be developed from a careful statistical analysis of public opinion. As an example, open access survey data such as the 'Wind Power Survey for Helsinki 2015' (Kaupunkiympäristön and Yleissuunnittelu 2016) combined with expert knowledge can guide the further extraction of the problem structure including the dynamics of reward emissions and state transitions.

We propose that AI methods are incorporated into mediating assessment protocols.

Additionally, AI methods can help practitioners and reporting teams to formally represent a problem. Specifically, natural language processing tools can be employed not only to gauge the sentiment of publicly accessible forums (e.g., social media or discussion boards) but also to identify key concepts and semantic correlations in large volumes of text. Therefore, these tools provide additional support in setting up a problem's state, action and reward spaces as well as its reward emission and state transi-

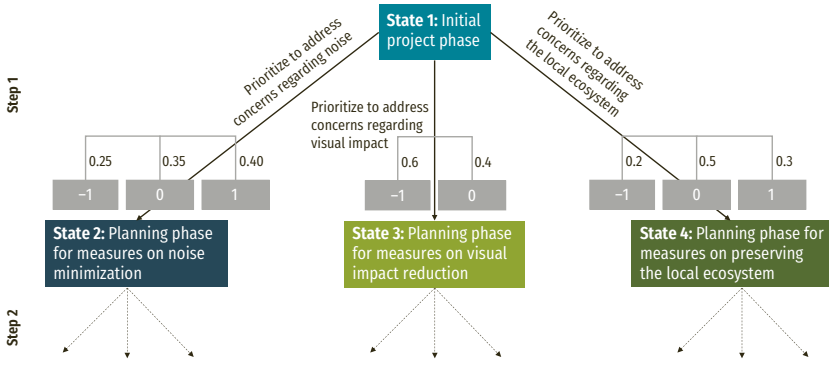
 https://doi.org/10.14512/tatup.32.1.43

tion dynamics, thereby making the problem amenable to optimization algorithms. For a comprehensive review of specific NLP algorithms, we refer to Jurafsky and Martin (2014).

Discussion

We have proposed a framework to aid policy decisions related to technology assessment. The MAP protocol is the basis for structuring relevant information and reaching consensus among decision-makers. MAP relies on detailed technical reports, established by expert staff, being distributed to decision-makers who then vote on various dimensions of a particular decision.

We propose that AI methods are incorporated into MAP. Combined with Bayesian methods and NLP, dynamic programming and reinforcement learning provide normative tools for finding optimal decisions subject to uncertainty. However, additional uncertainty may be introduced when establishing a formal representation of the decision-making problem, for instance if full expert knowledge is lacking, if potential biases are introduced through the choice of survey methods or statistical analyses thereof, or if additional systematic biases are introduced through human or automated data handling (e.g., using NLP methods to extract sentiments from text-based data sets). The link between MAP and the AI methods comprises expert practitioners that employ the AI methods and summarize the optimal action plans into reports that enter the MAP. While AI methods can be employed for assessing societal implications or risks of technology assessment, we focus here specifically on their potential role in improving decisions.



The gray tree structures represent hypothetical reward structures, that is the probabilities of particular public reactions in response to state-action pairs. In the interest of simplicity, we consider the case where only three reactions are recorded: negative sentiment (reward = −1), neutral (reward = 0), and positive sentiment (reward = 1). The corresponding probabilities could be obtained, for instance, by statistical analysis of relevant survey data. The black arrows map state-action pairs to new states. While we assume deterministic state transitions here, in principle these transitions could also be probabilistic.

For example, if in the initial phase concerns regarding the visual impact are prioritized, it would still be possible for an external factor such as the agenda of an environmental organization to sway the action towards a planning phase for measures addressing the preservation of the local ecosystem. Setting up such a problem structure requires substantial domain knowledge and could be supported by statistical analyses and NLP. Once a problem is formally represented, DP and RL algorithms can aid technology assessment by providing a framework to compute optimal sequences of actions.

Fig. 2: Schematic showing the problem structure pertaining to our example of the social acceptance of wind park construction. *Source: authors' own compilation*

To provide an application of our framework, we discussed the use of MAP in the context of the construction of a community wind park, and explored how AI methods could help to inform better policy measures seeking to improve the social acceptance of wind parks. Practical implementation of our framework requires the identification of relevant evaluation dimensions and the formal representation of key problems within an evaluation dimension.

To test the efficacy of our framework, we suggest an experimental design following Barham et al. (2014) and Holt and Laury (2002). By combining these standard methods for measuring risk and uncertainty in decisions with our proposed framework, the impact of AI methods on decision processes can be studied in a laboratory experiment. Since this experiment allows for the direct measurement of both risk and uncertainty, the anticipated reduction of these factors due to the incorporation of AI methods should become visible.

We conclude by addressing the limitations of our approach. It is possible that the MAP protocol cannot fully compensate for human bias in the decision-making process. Additionally, while reinforcement learning and dynamic programming algorithms adhere to a normative perspective, they are bounded by the formal representation of a problem, which, as noted above, is susceptible to bias. This poses an additional risk since humans may have particularly high levels of trust in machine-assisted decision-making processes. Furthermore, aspects of the protocols discussed here (e.g., the choice of advising experts or of the AI tools employed) could introduce path dependencies into the decision-making process that may affect decision outcomes. It is therefore important that our protocol be tested with regard to these or similar limitations (Katzenbach and Ulbricht 2019) in real applications or test setups in order to frame it within the larger debate of algorithmic policy making (Lenk 2018). Whether a comprehensive formal mapping of relevant technology assessment scenarios can be achieved is still to be explored; this is a pertinent question for future research.

References

Alesina, Alberto; Miano, Armando; Stantcheva, Stefanie (2020): The polarization of reality. In: AEA Papers and Proceedings 110, pp. 324–328. https://doi.org/10.1257/pandp.20201072

Barham, Bradford; Chavas, Jean-Paul; Fitz, Dylan; Rios-Salas, Vanessa; Schechter, Laura (2014): The roles of risk and ambiguity in technology adoption. In:

https://doi.org/10.14512/tatup.32.1.43

Journal of Economic Behavior & Organization 97, pp. 204–218. https://doi.org/10.1016/j.jebo.2013.06.014

Beiderbeck, Daniel; Frevel, Nicolas; von der Gracht, Heiko; Schmidt, Sascha; Schweitzer, Vera (2021): Preparing, conducting, and analyzing Delphi surveys. Cross-disciplinary practices, new directions, and advancements. In: MethodsX 8, p. 1–20. https://doi.org/10.1016/j.mex.2021.101401

Bellman, Richard (2010): Dynamic programming. With a new introduction by Stuart Dreyfus. Princeton: Princeton University Press.

Bertsekas, Dimitri (2019): Reinforcement learning and optimal control. Belmont: Athena Scientific.

Bertsekas, Dimitri; Tsitsiklis, John (1996): Neuro-dynamic programming. Belmont: Athena Scientific.

Holt, Charles; Laury, Susan (2002): Risk aversion and incentive effects. In: American Economic Review 92 (5), pp. 1644–1655. https://doi.org/10.1257/000282802762024700

Jurafsky, Dan; Martin, James (2014): Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Pearson.

Kahneman, Daniel (2011): Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Kahneman, Daniel; Lovallo, Dan; Sibony, Olivier (2019): A structured approach to strategic decisions. In: MIT Sloan Management Review 60 (3), 04. 03. 2019, pp. 67–73. Available online at https://sloanreview.mit.edu/media-download/65293/a-structured-approach-to-strategic-decisions/, last accessed on 06. 02. 2023.

Kahneman, Daniel; Sibony, Olivier; Sunstein, Cass (2021): Noise. A flaw in human judgment. New York: Little, Brown Spark.

Katzenbach, Christian; Ulbricht, Lena (2019): Algorithmic governance. In: Internet Policy Review 8 (4), pp. 1–18. https://doi.org/10.14763/2019.4.1424

Kaupunkiympäristön, Helsingin; Yleissuunnittelu, Maankäytöön (2016): Wind power survey for Helsinki 2015, 23. 08. 2016. Available online at https://hri.fi/data/en_GB/dataset/helsingin-tuulivoimakysely-2015, last accessed on 06. 02. 2023.

Leiren, Merethe; Aakre, Stine; Linnerud, Kristin; Julsrud, Tom; Di Nucci, Maria-Rosaria; Krug, Michael (2020): Community acceptance of wind energy developments. Experience from wind energy scarce regions in Europe. In: Sustainability 12 (5), pp. 1–22. https://doi.org/10.3390/su12051754

Lenk, Klaus (2018): Formen und Folgen algorithmischer Public Governance. In: Resa Mohabbat Kar, Basanta Thapa and Peter Parycek (eds.): (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft, pp. 228–267. Berlin: Kompetenzzentrum Öffentliche IT. Available online at https://www.oeffentliche-it.de/documents/10181/14412/(Un)berechenbar+-+Algorithmen+und+Automatisierung+in+Staat+und+Gesellschaft, last accessed on 06. 02. 2023.

Russell, Stuart; Norvig, Peter (2021): Artificial intelligence: A modern approach. Harlow: Pearson.

Stolwijk, Sjoerd; Vis, Barbara (2020): Politicians, the representativeness heuristic and decision-making biases. In: Political Behavior 43 (4), pp. 1411–1432. https://doi.org/10.1007/s11109-020-09594-6

Sutton, Richard; Barto, Andrew (2018): Reinforcement learning. An introduction. Cambridge: MIT Press.

Wiering, Marco; van Otterlo, Martijn (eds.) (2012): Reinforcement learning. State-of-the-art. Softcover reprint. Heidelberg: Springer. https://doi.org/10.1007/978-3-642-27645-3

DR. LILLA HORVATH
completed her PhD in computational cognitive science at Free University of Berlin with a one-year research visit at New York University. She now works in the Science Group at PD – Berater der öffentlichen Hand GmbH as a public sector consultant focusing on AI-related topics.

DR. ERICH RENZ
holds a PhD in economics from the University of Regensburg and worked as a Science Group Senior Consultant at PD – Berater der öffentlichen Hand GmbH in the field of data analytics. In his research, he conducts online, laboratory, and field experiments in the areas of organizational change, entrepreneurial behavior, and innovation decision processes.

DR. CHRISTIAN ROHWER
is a senior consultant at PD – Berater der öffentlichen Hand GmbH. He works on projects in the Science Group with a focus on artificial intelligence. Previously, he was a researcher at the Max Planck Institute for Intelligent Systems in Stuttgart.

DR. DANIEL SCHURY
is as senior consultant in the Science Group of PD – Berater der öffentlichen Hand GmbH, where he focuses his work on data projects in federal ministries. He studied atomic physics at JLU Gießen before moving to the GSI Helmholtzzentrum in Darmstadt were he pursued his PhD, followed by two postdoctorial stays in Paris and New York.

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom

RESEARCH ARTICLE

The politics of models: Socio-political discourses in modeling of energy transition and transnational trade policies

Titus Udrea*,[1], Leo Capari[2], Anja Bauer[3]

Abstract • In this article, we discuss the (re)production of socio-political discourses in two modeling communities, energy transition and transnational trade. Methodologically, we build on bibliometric and qualitative analyses of academic articles. Our analyses show how models structure epistemic communities and are closely linked to specific discourses. The modeling of the energy transition is driven by and contributes to discourses on mitigating climate change and access to energy. Different trade models address either multilateral or regional trade, yet in each case favoring international trade. Overall, we illustrate how the 'politics of models' does not only concern their use at the science-policy interface, but is already inscribed in their development, application, and scientific exploitation. These analyses may help experts, policy makers, and the public to better assess the knowledge claims and evidence politics of computer modeling.

***Die Politik der Modelle**: Soziopolitische Diskurse in der Modellierung der Energiewende und transnationaler Handelspolitiken*

Zusammenfassung • *Wir untersuchen die (Re-)Produktion soziopolitischer Diskurse in zwei Modellierungsgemeinschaften, der Energiewende und dem transnationalen Handel. Methodisch stützen wir uns auf bibliometrische und qualitative Analysen akademischer Artikel. Unsere Analysen zeigen, wie Modelle epistemische Gemeinschaften strukturieren und eng mit spezifischen Diskursen verbunden sind. Die Modellierung der Energiewende wird von Diskursen über die Eindämmung des Klimawandels sowie dem Zugang zu Energie angetrieben und prägt diese. Verschiedene Handelsmodelle befassen sich entweder mit multilateralem oder regionalem Handel. Wir zeigen somit, dass die ‚Politik der Modelle' nicht erst bei der Verwendung an der Schnittstelle zwischen Wissenschaft und Politik sichtbar wird, sondern bereits in ihre Anwendung und Verwertung eingeschrieben ist. Diese Analysen können Expert*innen, Entscheidungsträger*innen und der Öffentlichkeit helfen, die Wissensansprüche der Computermodellierung besser zu beurteilen.*

Keywords • *policy modeling, discourses, energy transition, trade policy*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Introduction

Energy policy and transnational trade policy are two main areas of 'modeling for policy'. Both areas are highly relevant for the economic performance of countries, leading to increasing demand for 'usable' knowledge about the effects of international trade or the energy provision of a country. In these contexts, computational models fulfil various functions, from identifying and analyzing societal problems to examining different policy instruments (e.g., energy mixes and energy savings) and assessing the impacts and costs of planned and implemented policies. They also serve to justify and legitimize public action, e.g., when models legitimize free trade agreements based on projected growth scenarios. In this role, models are not neutral tools, providing orientation and answers to (exogenously given) societal or political questions, but models are simplifications that

* *Corresponding author: titus.udrea@oeaw.ac.at*
[1] *Institute of Technology Assessment, Austrian Academy of Sciences, Vienna, AT*
[2] *Vienna Doctoral School of Ecology and Evolution, University of Vienna, Vienna, AT*
[3] *Institut für Technik- und Wissenschaftsforschung, Universität Klagenfurt, Klagenfurt, AT*

https://doi.org/10.14512/tatup.32.1.49
Received: 26. 08. 2022; revised version accepted: 17. 01. 2022;
published online: 23. 03. 2023 (peer review)

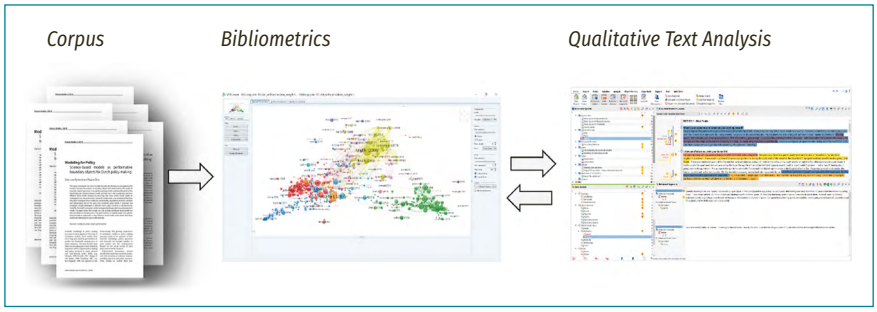


Fig. 1: Research design. *Source: authors' own compilation*

embed and are embedded in specific ideas, assumptions, paradigms, framings, discourses, or representations.

Countering the common notion of computational models as neutral representations, social science scholars have conceived of models as 'techniques of futuring' (Hajer and Pelzer 2018; Oomen et al. 2022), 'boundary objects' (van der Heide 2020; Zeiss and van Egmond 2014) or 'performative devices' (Svetlova 2012). Such conceptions emphasize the social role of models in coordinating different actors (including modelers, other scientists, policy-makers and stakeholders), in contributing to shared understandings and imaginaries about socio-political issues and futures, and in rendering these understandings and imaginaries into 'authoritative orientations for action' (Oomen et al. 2022; van der Heide 2020; Zeiss and van Egmond 2014). Hence, models do not simply inform policy-making; they have performative power by making and shaping the socio-political discourses and worlds in which they are embedded (van der Heide 2020; Zeiss and van Egmond 2014). Against this background, the article explores how scientific computer-based modeling is embedded in and (re)-produces socio-political discourses in the two domains, energy transition and transnational trade. Following Hajer (1997, p. 44), a discourse is a "specific ensemble of ideas, concepts, and categorizations that are produced, reproduced, and transformed in a particular set of practices" (in our case modeling) and which give meaning to social or physical phenomena.

The analysis suggests that energy modeling is strongly connected to two main discourses, the mitigation of global climate change and energy access and security.

Research design

Models coordinate actors and shape discourses in different sites, inter alia, policy processes, scientific debates, or media communication. In this article, we focus on the socio-political discourses that are reproduced in scientific modeling communities by analyzing the respective scientific literature in the two domains. Of course, scientific articles alone do not produce socio-political discourses; they are first and foremost communications within the scientific community. However, by analyzing scientific modeling literature, we seek to trace how socio-political narratives become dressed in techniques and how policy options become potentially narrowed down by rendering certain questions and "trajectories of action possible and maybe even logical" while excluding others (van der Heide 2020, p. 125).

In the following, we give an overview of the main steps of our research design that combines bibliometric and qualitative text analyses (Figure 1). For more detailed accounts, please consult the supplementary research data material.

First, using the Scopus database, we constructed a corpus for each domain, consisting of scientific journal articles that deal with computational modeling. We then applied several bibliometric analyses to each corpus individually. The Scopus analysis tool provided a first overview of the corpora, including publications per year, sources of publication, geographic and institutional hotspots, and authorship frequency. For our main analysis, we used the software tool VOSviewer, which facilitates the creation and visualization of various bibliometric networks. Bibliometric networks consist of nodes representing authors, articles, publications, or key terms and their relations (van Eck and Waltman 2014). We explored several bibliometric relations (term co-occurrences, co-citation, and bibliographic coupling), each providing a distinct way to structure and contextualize the bibliographic information of the corpora (see supplementary 'Research Data'). For the present discussion, however, we draw only on the citation networks. While not presented in this paper, complementary bibliometric analyses validated our interpretation of the citation networks. The citation network identifies key corpus publications (number of citations) and article clusters (articles linked by citations). We adjusted the clustering parameters and resolution, following van Eck and Waltman (2013), to fit the research question and the purpose of providing broad examples within the two domains. Clusters of linked articles indicate similar research foci; however, clusters are not rigidly separated and isolated, but their borders are fuzzy and are connected by intermediating au-

 https://doi.org/10.14512/tatup.32.1.49

thors and sub-clusters. In this article, we focus only on the main clusters as an exemplary basis for the qualitative analysis of the respective discourses.

The citation network alone does not reveal information about the modeling approaches or discourses. Therefore, we first characterized clusters by screening the titles, abstracts, and key-words of core publications. We particularly looked for the stated modeling approaches and models as well as the themes, topics and problems addressed in the titles and abstracts. Secondly, to enrich our discussion of the socio-political discourses, we selected 20 key articles from each domain for an in-depth qualitative analysis. We analyzed the texts along dimensions such as modeling approach, socio-political contextualization, and policy recommendations. For the purpose of the present discussion, we reconstructed the dominant problem and solution narratives in the selected articles. Problem narratives refer to the construction of the underlying problem or question for which the modeling activities are applied and its embedding in wider socio-political issues and debates. Solution narratives consist of those statements that link the modeling results to (actual or potential) socio-political actions.

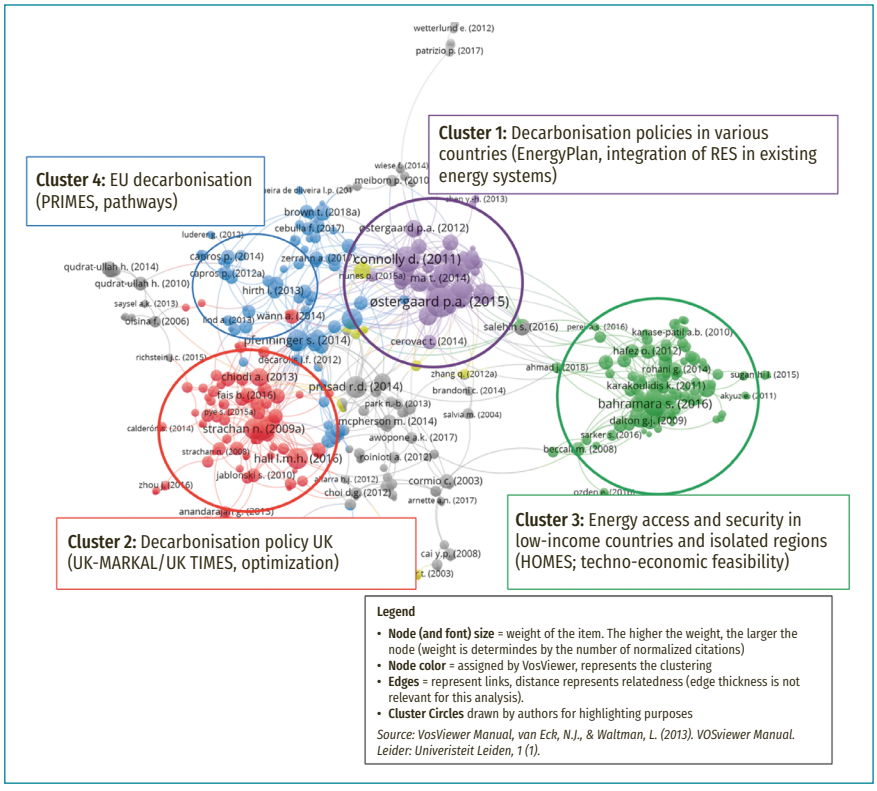


Fig. 2: Energy transition modeling: core citation network. *Source: authors' own compilation*

Modeling the energy transition: between combatting climate change and ensuring energy security

Energy system models have been entangled with energy policies since the 1950s and 1960s to calculate countries' energy demand and respective energy provisions. In the last decades, models have increasingly been developed to address the transition of energy systems from fossil fuels to renewable or hybrid energy sources. Our corpus on modeling the energy transition includes 1.242 articles (1974–2018), with a steady increase in publications from the early 2000s. Figure 2 shows the core citation map[1] with the most cited and linked articles grouped into clusters. We restrict the following discussion to the four most relevant and distinct clusters – in terms of number of articles, citations, links, as well as modeling approaches and socio-political issues.

The citation map provides first insights into the modeling community: Cluster 1 (C1, violet) mainly focuses on the model EnergyPLAN in different geographic contexts (e.g., Ireland, Denmark, Hong Kong) to simulate the potential for the transformation to renewable energy systems. Cluster 2 (C2, red) mainly revolves around the UK energy system decarbonization and discusses MARKAL, times, and hybrid modeling approaches. In Cluster 3 (C3, green), the model HOMER is applied to optimize hybrid renewable energy systems in rural regions, less industrialized countries, and emerging economies. Cluster 4 (C4, blue) focuses on PRIMES to explore decarbonization pathways of the EU.

We observe a dominance of a few modeling tools with community structuring effects, i.e. citation networks built around them. Most importantly, the analysis suggests that energy modeling is strongly connected to two main discourses, the mitigation of global climate change and energy access and security. The discourse on climate change mitigation is dominant (C1, C2, C4). The transition of regional and national energy systems from fossil fuels to renewable energies is presented as a key solution to decarbonize economies and reduce green house gas (GHG) emissions. This societal relevance is not only invoked abstractly, but modeling is deeply contextualized in specific policies, aims or strategies, such as the EU climate targets (C4), the UK decarbonization strategy (C2), or, as the following quote illustrates Hong Kong's climate ambitions:

> „Climate change, the defining challenge of our time, brings about more severe weather and affects each and every one

1 Core citation map includes only those articles with at least 10 citations and further excludes isolates. The core citation map contains 410 articles.

> of us. […] Hong Kong has set a target to achieve a reduction in energy intensity of at least 25 % by 2030 corresponding to the APEC (Asia-Pacific Economic Cooperation) target and a reduction of carbon intensity by 50–60 % by 2020 corresponding to the Copenhagen Accord […]" (Ma et al. 2014, p. 301).

Such policy targets not only serve as a motivation for modeling but are central reference points to model respective trends, strategies, or scenarios. Modeling provides knowledge on strategies and measures to reach these targets (e.g., the optimal energy mix), the associated costs, uncertainties, sensitivities, and challenges (e.g., in terms of flexibility). Optimization models like TIMES or MARKAL (C2) calculate the optimal technology mix to meet the energy demand at minimum costs. Other models, such as EnergyPLan (C1), explore the potential for national renewable energy system (RES) transformation and respective CO_2 reductions by simulating and comparing energy scenarios. They also explore the integration and optimal combinations of renewable energy sources into the existing electricity or transport sectors.

In most cases, modeling results are intended to inform policies, yet with varying roles, ranging from general orientation and insights into developments and uncertainties, to supporting or challenging policies with concrete recommendations. In the following illustrative quote, the modeling results support UK's decarbonization strategy:

> "For UK policy makers an important message from this modelling exercise is that deep long-term CO_2 emission reductions are feasible" (Strachan and Kannan 2008, p. 2961).

The authors further highlight the role of Carbon Capture and Storage as a critical zero-carbon technology, thereby influencing the narrative on potential solutions. In other cases, modelers challenge government policies:

> "[…] However, it is obvious that the RE scenario is much better than the governmentally proposed scenario in terms of climate change, availability, and long term sustainability" (Ma et al. 2014, p. 307).

Hence, modeling not only draws on the common problem narrative of climate change but actively contributes to solution narratives by supporting or challenging policies. The second socio-political discourse revolves around energy access and security (C3). The stated problem is that many people in rural, low-income countries or geographically isolated regions have little or no access to electricity, hindering their economic and social development:

> "[…] access to electricity is fundamental in advancing the well-being of any society and promoting economic growth […] Like most countries in sub-Saharan Africa, access to electricity in Ghana is low compared to more advanced countries" (Adaramola et al. 2014, p. 284).

In this discourse, the transition is a solution not primarily because GHG emissions can be reduced but because RES are cost-effective, decentralized, and stable and can provide energy access to energy-poor people and regions.

> "In isolated areas, the high cost of transmission lines and higher transmission losses are encouraging the use of green sources of energy. Combining two or more renewable energy sources […] gives a stable energy supply in comparison to non-renewable energy systems" (Hafez and Bhattacharya 2012, p. 7).

In these contexts, modeling provides insights into the technical and economic feasibility of hybrid renewable energy systems in specific geographical contexts (countries, regions, islands). The dominating modeling tool is HOMER, which, in contrast to the optimization models introduced above, does not primarily focus on CO_2 reductions but cost-effective energy provision and security. Therefore, the modeling may also result in solutions that combine fossil fuel and renewable energy sources:

> "Therefore, any government policy that aid decreasing the cost of diesel fuel can reduce the operating cost of hybrid power system in remote and semi-urban area of the country" (Adaramola et al. 2014, p. 291).

In sum, energy system transition and respective modeling activities are motivated by two underlying problems: climate change and energy poverty. As a result, modeling activities have different foci, particularly apparent in optimization models, searching for cost-effective solutions to reduce a certain amount of GHG or secure a certain level of energy security. The discourses also have clear geographic foci. While the climate change discourse is strongly linked to industrialized countries, the energy security discourse is linked to less industrialized countries and rural areas.

Trade modeling: between favoring bilateral and multilateral trade agreements

Trade policy was among the first domains to use computational models. Modeling entered mainstream policy-making in the 1970s at the UN, as a way to scientifically and empirically validate macroeconomic policies (Estrada and Park 2018). Since then, "macroeconomic modelling has become an almost mandatory part of assessing the impacts of a new (trade) policy" (Pollitt 2018). Our corpus (1834 documents, 1966–2018) reflects this, showing a steady increase in publications since the late 1980s.

 https://doi.org/10.14512/tatup.32.1.49

Figure 3 shows the core citation network[2] of the trade modeling corpus. For the purposes of this paper, the largest two clusters are considered as the basis for the further qualitative analysis. The first cluster (blue) revolves around computable general equilibrium (CGE) modeling, multilateral trade agreements and trade effects-related topics. The second cluster (red) is dominated by gravity models and research topics such as regional markets, bilateral and domestic trade effects, welfare and labor. CGE and gravity models, as well as the underlying economic schools of thought, have been used for trade policy analyses for over 50 years (Piermartini and Teh 2005).

While both clusters are firmly embedded in the mainstream free trade discourse (i.e., positive framing of free trade as central for welfare), the two clusters and respective modeling approaches include different foci and relate to distinct policies, i.e., multilateral versus regional free trade.

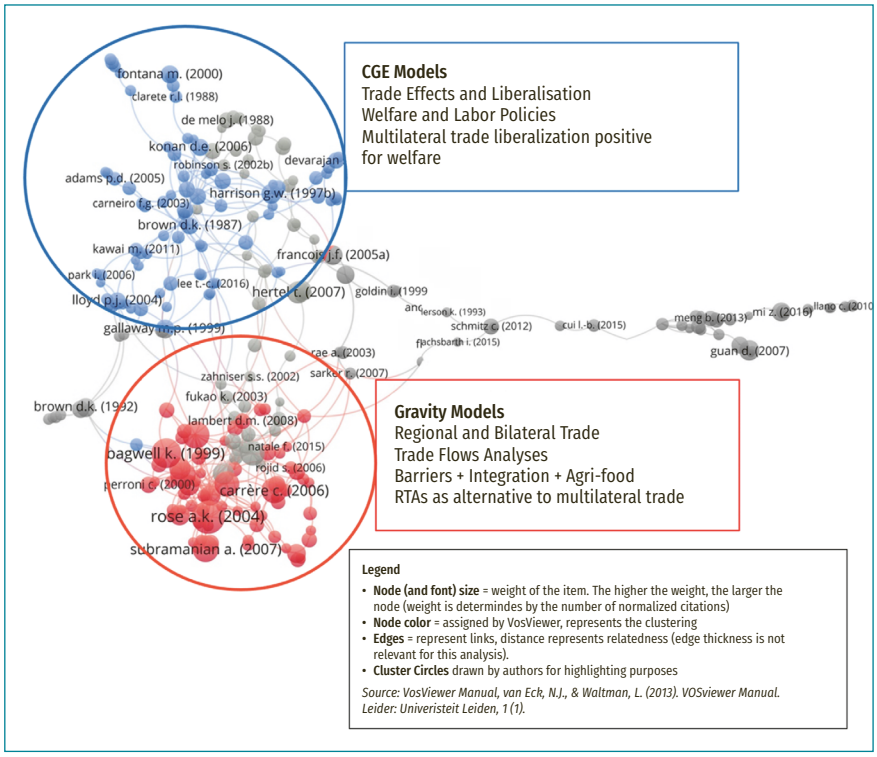


Fig. 3: Trade core citation network and focus clusters. *Source: authors' own compilation*

The first cluster is dominated by CGE models, which calculate economy-wide impacts and welfare effects of trade liberalization across sectors, either ex-post (after the agreement) or ex-ante (before the agreement). Modeled welfare effects include, e.g., gross domestic product, growth or sector-specific effects, such as wages and labor dynamics. CGE models by their design (built on mainstream economic theory that trade liberalization, in the long run, always increases welfare) provide a generally positive outlook of trade arrangements. With a broad economic scope, CGE-based analyses tend to conclude (as well formulate policy recommendations) that multilateral trade arrangements are the most conducive to welfare increases. Multilateral trade refers to the global trade system and broad international frameworks, e.g., the World Trade Organization (WTO) with more than 150 members. Several studies in our sample specifically compare regional and multilateral trade arrangements and recommend the former as the better policy alternative:

> "[…] while regional and bilateral FTAs may be welfare enhancing for the member countries directly involved, these welfare gains are considerably smaller than those resulting from multilateral trade liberalization, and, in any case, accrue in absolute terms primarily to the large industrialised countries" (Brown et al. 2003, p. 826).

Turning to the second cluster, where, along gravity models, topics involve regional trade agreements and regional economic integration, e.g., The North American Free Trade Agreement (NAFTA), The Association of Southeast Asian Nations (ASEAN), African Union. Regional trade arrangements are also compared to multilateral arrangements (e.g., WTO) in terms of trade flows. Bilateral agreements are also included here, and new large regional trade frameworks, such as the proposed TTIP (Transatlantic Trade and Investment Partnership) between the US and EU, or the 16-member TPP (Trans-Pacific Partnership), which blur the lines between multilateral and regional trade. There is a stronger geographic focus evidenced by the presence of regional items and specific countries, such as 'Middle East', 'South Asia' (e.g., Bangladesh), 'East Asia', 'Africa' (e.g., Tunisia, Morocco), 'South and East Africa'.

With a focus on spatial interaction, gravity models investigate regional trade flows and the effects of trade agreements on economic growth. Usually employed in ex-post analyses, gravity models are used to assess the role of trade deals for trade and economic growth:

> "During the last 10 years, regionalism has re-emerged as a major issue in the policy agenda [listing several agreements; the authors …]. The effect of this 'second wave' of regionalism on trade is still an open question. Do they really increase trade among members? Do they contribute to further trade liberalization with nonmember countries or

2 Core citation map includes only those articles with at least 10 citations and excludes isolates. The core citation map contains 253 articles.

> undermine it? Do they harm nonmember countries?" (Soloaga and Winters 2001, p. 1).

Often, the focus is on specific sectors, particularly agriculture (as a function of interregional trade effects). Newer literature, using gravity modeling, generally shows positive effects of regional trade and presents bilateral, regional trade as an alternative to multilateral trade. Especially for the agricultural sector, regional trade is seen as a more pragmatic approach to trade liberalization: "RTAs may be an attractive alternative to promote agricultural trade. […] While the GATT/WTO has not made significant progress in agriculture, regional trade agreements offer countries the opportunity to liberalize and increase members' agricultural trade. This may be a fundamental explanation for the aggressive pursuit of regionalism over the past two decades" (Grant and Lambert 2008, p. 765).

Regarding policy recommendations, the takeaway is that regional trade is the alternative because multilateral arrangements are complicated to establish and maintain (yet not because the welfare effects would be higher).

Energy system transition and respective modeling activities are motivated by two underlying problems: climate change and energy poverty.

Conclusions

In this article, we have illustrated how modeling approaches are entangled with specific problem framings, analytical foci, and geographic regions, and support different policies: Modeling energy system transitions refers to the climate change discourse but also to energy access and security discourses in poor energy access regions. The two dominant trade modeling approaches refer to distinct policies, i.e., multilateral versus regional free trade. A limitation of this contribution is the cursory (and exemplary) presentation of the various bibliometric and qualitative analyses due to word limits. Nevertheless, our analyses offer valuable insights on the socio-political nature of models. Analyzing how models discursively present policy problems and solutions exemplifies their roles as 'technologies of futures' or 'boundary objects' and how they might reinforce, consolidate or transform the possible repertoires for action (Oomen et al. 2022). Broader socio-political aims and specific policies frequently serve as the motivational background of modeling and directly feed into modeling, e.g., by serving as endpoints for which optimized solutions are sought. Moreover, the different policy discourses feature a geographic component in both modeling domains. The geographic foci include a strong political component linked to questions of welfare, poverty and equality (trade) or energy security (energy), respectively. This raises questions about the political implications of this geographic imbalance, particularly when considering that less-industrialized countries are frequently the object of modeling in energy transition and trade policy, while respective models have been developed in the Global North. In line with previous research, we thus observe how "agendas, commitments, goals [and strategies] of policy communities are referred to in scientific communities as justifications" for modeling approaches and how, in return, these scientific commitments might cement the policy goals that motivated them initially (Demortain 2017, p. 147; Shackley and Wynne 1996, p. 221).

Overall, with our analysis of scientific modeling discourses, we illustrate how the 'politics of models' does not only concern their use at the science-policy interface; modeling communities frequently make "authoritative claim[s] to policy-relevant knowledge" (Knorr-Cetina 2007, p. 365) already in the modeling process. The insights may serve both TA academics and policy experts to assess knowledge claims and epistemic communities. It is not only technical questions regarding the modeling algorithms, data, validity, uncertainties etc., that are of relevance for their use and assessment but also their discursive embedding and power. TA has long been active in analyzing political and societal debates and discourses on technologies, including respective visions and expectations.

Funding • This article is based on research conducted in the project 'CoMoPA – Computational Modelling for Policy Advice' funded by the Austrian Academy of Science Innovation Fund 'Research, Science and Society' [no. IF_2017_13]

Competing interests • The authors declare no competing interests.

References

Adaramola, Muyiwa; Agelin-Chaab, Martin; Paul, Samuel (2014): Analysis of hybrid energy systems for application in southern Ghana. In: Energy Conversion and Management 88, pp. 284–295. https://doi.org/10.1016/j.enconman.2014.08.029

Brown, Drusilla; Deardorff, Alan; Stern, Robert (2003): Multilateral, regional and bilateral trade-policy options for the United States and Japan. In: World Economy 26 (6), pp. 803–828. https://doi.org/10.1111/1467-9701.00549

Demortain, David (2017): Expertise, regulatory science and the evaluation of technology and risk. Introduction to the special issue. In: Minerva 55 (2), pp. 139–159. https://doi.org/10.1111/1467-9701.00549

Estrada, Mario; Park, Donghyun (2018): The past, present and future of policy modeling. In: Journal of Policy Modeling 40 (1), pp. 1–15. https://doi.org/10.1016/j.jpolmod.2018.01.003

Grant, Jason; Lambert, Dayton (2008): Do regional trade agreements increase members' agricultural trade? In: American Journal of Agricultural Economics 90 (3), pp. 765–782. https://doi.org/10.1111/j.1467-8276.2008.01134.x

Hafez, Omar; Bhattacharya, Kankar (2012): Optimal planning and design of a renewable energy based supply system for microgrids. In: Renewable Energy 45, pp. 7–15. https://doi.org/10.1016/j.renene.2012.01.087

Hajer, Maarten; Pelzer, Peter (2018): 2050 – An energetic odyssey. Understanding 'techniques of futuring' in the transition towards renewable energy. In: Energy Research & Social Science 44, pp. 222–231. https://doi.org/10.1016/j.erss.2018.01.013

Hajer, Maarten (1997): The politics of environmental discourse: Ecological modernization and the policy process. Oxford: Oxford University Press. https://doi.org/10.1093/019829333x.001.0001

Knorr-Cetina, Karin (2007): Culture in global knowledge societies. Knowledge cultures and epistemic cultures. In: Interdisciplinary Science Reviews 32 (4), pp. 361–375. https://doi.org/10.1179/030801807X163571

Ma, Tao; Østergaard, Poul; Lund, Henrik; Yang, Hongxing; Lu, Lin (2014): An energy system model for Hong Kong in 2020. In: Energy 68, pp. 301–310. https://doi.org/10.1016/j.energy.2014.02.096

Oomen, Jeroen; Hoffman, Jesse; Hajer, Maarten (2022): Techniques of futuring. On how imagined futures become socially performative. In: European Journal of Social Theory 25 (2), pp. 252–270. https://doi.org/10.1177/13684310209888826

Piermartini, Roberta; Teh, Robert (2005): Demystifying modelling methods for trade policy. WTO Discussion Paper, No. 10. Available online at https://www.econstor.eu/bitstream/10419/107045/1/wto-discussion-paper_10.pdf, last accessed on 04. 02. 2023.

Pollitt, Hector (2018): What is macroeconomic modelling? And why do we do it? Available online at https://www.camecon.com/blog/what-is-macroeconomic-modelling-and-why-do-it/, last accessed on 04. 02. 2023.

Shackley, Simon; Wynne, Brian (1996): Representing uncertainty in global climate change science and policy. Boundary-ordering devices and authority. In: Science, Technology, & Human Values 21 (3), pp. 275–302. https://doi.org/10.1177/016224399602100302

Soloaga, Isidro; Winters, Alan (2001): Regionalism in the nineties. What effect on trade? In: The North American Journal of Economics and Finance 12 (1), pp. 1–29. https://doi.org/10.1016/S1062-9408(01)00042-0

Strachan, Neil; Kannan, Ramachandran (2008): Hybrid modelling of long-term carbon reduction scenarios for the UK. In: Energy Economics 30 (6), pp. 2947–2963. https://doi.org/https://doi.org/10.1016/j.eneco.2008.04.009

Svetlova, Ekaterina (2012): On the performative power of financial models. In: Economy and Society 41 (3), pp. 418–434. https://doi.org/10.1080/03085147.2011.616145

van der Heide, Arjen (2020): Model migration and rough edges. British actuaries and the ontologies of modelling. In: Social Studies of Science 50 (1), pp. 121–144. https://doi.org/10.1177/0306312719893465

van Eck, Nees; Waltman, Ludo (2013): VOSviewer manual. Available online at https://www.vosviewer.com/documentation/Manual_VOSviewer_1.5.4.pdf, last accessed on 04. 02. 2023.

van Eck, Nees; Waltman, Ludo (2014): Visualizing bibliometric networks. In: Ying Ding, Ronald Rousseau and Dietmar Wolfram (eds.): Measuring scholarly impact. Methods and practice. Cham: Springer, pp. 285–320. https://doi.org/10.1007/978-3-319-10377-8_13

Zeiss, Ragna; van Egmond, Stans (2014): Dissolving decision making? Models and their roles in decision-making processes and policy at large. In: Science in Context 27 (4), pp. 631–657. https://doi.org/10.1017/S0269889714000234

Research Data

https://epub.oeaw.ac.at/ita/ita-papers/the_politics_of_models_supplementary.pdf

TITUS UDREA
is academy scientist at the Institute of Technology Assessment, Austrian Academy of Sciences. He has a background in political science and currently researches in the areas of policy modeling, AI governance, and energy innovation.

LEO CAPARI
is human ecologist, currently PhD candidate at the University of Vienna and was a junior scientist at the Institute of Technology Assessment, Austrian Academy of Sciences. His main research and methodological interests are sustainability sciences and computer-based social network analysis and visualization.

ASS.-PROF. DR. ANJA BAUER
is assistant professor at the Department of Science, Technology and Society Studies at the University of Klagenfurt. She researches and teaches in the areas of environmental, sustainability and technology governance with a special interest in the role of expertise, anticipation and participation in policy-making.

RESEARCH ARTICLE

Modeling sustainable mobility: Impact assessment of policy measures

Johannes Weyer*[1,2], Fabian Adelt[2], Marlon Philipp[2]

Abstract • Sociologically based models of complex systems can help to estimate the impact of policy measures on individuals and explain the resulting system dynamics. Using the example of the Ruhr region and the mobility of the people living there, the article demonstrates the concept of agent-based modeling, which draws on assumptions from analytical sociology and distinguishes between different types of actors. Simulation experiments conducted as part of the InnaMoRuhr project show significant differences in the behavior of these types, especially in their response to policy interventions. Policymakers should take this into account when planning and designing measures aimed at sustainable transformation.

Modellierung nachhaltiger Mobilität:
Eine Abschätzung der Wirkung politischer Maßnahmen

Zusammenfassung • *Soziologisch fundierte Modelle komplexer Systeme können dazu beitragen, die Wirkung politischer Maßnahmen auf einzelne Individuen abzuschätzen und die daraus resultierenden Systemdynamiken zu erklären. Am Beispiel des Ruhrgebiets und der Mobilität der dort lebenden Menschen demonstriert der Beitrag das Konzept einer agentenbasierten Modellierung, die auf Annahmen der analytischen Soziologie rekurriert und insbesondere verschiedene Akteurtypen unterscheidet. Simulationsexperimente, die im Rahmen des Projekts InnaMoRuhr durchgeführt wurden, zeigen erhebliche Unterschiede im Verhalten dieser Typen, insbesondere in ihrer Reaktion auf politische Interventionen. Politik sollte dies bei der Planung und Konzeption von Maßnahmen berücksichtigen, deren Ziel die nachhaltige Transformation ist.*

* *Corresponding author: johannes.weyer@tu-dortmund.de*
[1] *Senior professorship for sustainable mobility, TU Dortmund University, Dortmund, DE*
[2] *Social Research Centre, TU Dortmund University, Dortmund, DE*

https://doi.org/10.14512/tatup.32.1.56
Received: 20. 08. 2022; revised version accepted: 24. 11. 2022;
published online: 23. 03. 2023 (peer review)

Keywords • *analytical sociology, agent-based modeling, policy assessment, transportation*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Introduction: ABM and policy assessment

Policymakers in areas like transportation, energy, climate, or health, who are planning to introduce new regulations, are usually guided by expectations about the impact of these measures. For example, restrictions during the COVID-19 pandemic should slow down the spread of the virus, and the reductions of public transport prices were aimed at changing modal shift as well as lowering CO_2 emissions.

Frequently, expert advice has been used to assess the potential impact of policy measures in advance and to discuss and evaluate alternative strategies, e.g., in the case of COVID-19 vaccination. Typically, system dynamics models such as SEIRD are used to depict complex interactions between various factors: the numbers of susceptible (S), exposed (E), infectious (I), recovered (R) and dead (D) persons. The online COVID-19 Simulator 'CoSim' (Dings et al. 2022) connects these variables by functions and rates, such as the reproduction rate. This kind of mathematical modeling of system dynamics (SD) allows to adjust various parameters, e.g., the booster willingness, and to assess the impact of those measures by means of simulation experiments (Figure 1). Referring to CoSim, policy makers can assess prospectively, e.g., the effects of reducing the number of quarantine days, thus looking into the future, which can hardly be achieved with other methods.

However, SD models such as CoSim do not – or only partially – consider peoples' individual behavior and its impact on system dynamics. As can be seen in Figure 1, booster willingness is a global variable, applying equally to every person.

Agent-based modeling (ABM) has been established as an alternative approach with some advantages, but disadvantages as well. ABM takes into account the heterogeneity of real people, e.g., in terms of booster willingness, and conceives the 'social elements' of the system as individuals with subjective preferences. It takes a closer look at (i) the individual decision-making process, (ii) the factors that might, or might fail to, influence decision-making, and, finally, (iii) the systemic effects that result from the complex interactions of many people acting in different ways.

ABM allows focusing on typical patterns of agents' behavior, that must be addressed differently by policy measures. For example, lowering prices of public transport will not change mobility patterns of people who are not price sensitive and who are used to taking the car without considering alternatives. SD models might fail to grasp these differences between various agent types. Although ABM is more realistic in this respect, it also has some disadvantages. Due to the huge computer performance needed to calculate each individual decision, simulation experiments can only be conducted with small populations of some 10,000 agents, not with the population of an entire country. Thus, we suggest that combining the benefits of SD (i.e., global population) and ABM (i.e. individual behaviors) might be useful, to better support politicians preparing difficult decisions in the field of transportation, energy, climate, or health policies.

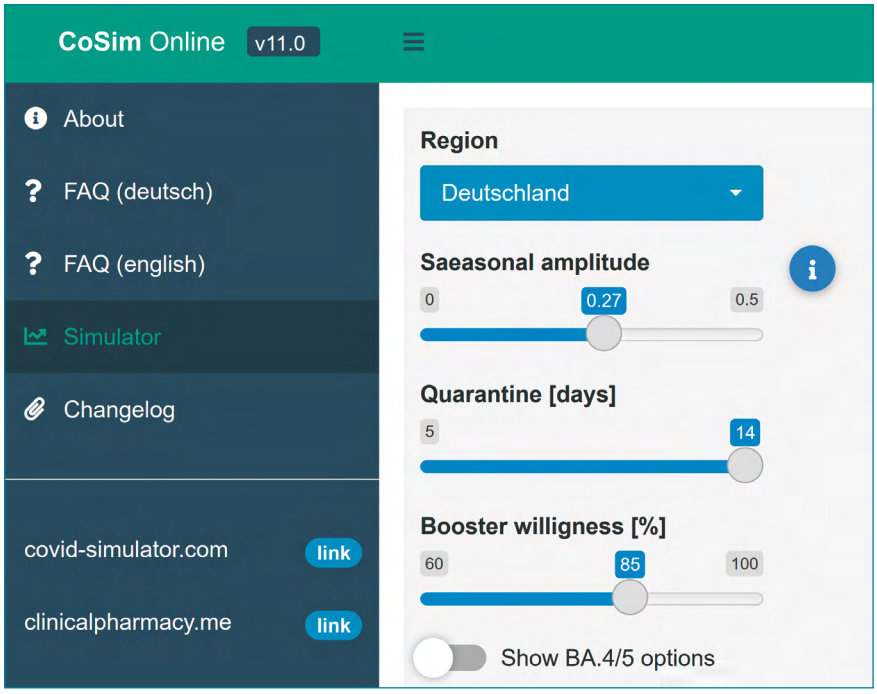


Fig. 1: CoSim COVID-19 Simulator. *Source: https://shiny.covid-simulator.com/covidsim*

Modeling complex socio-technical systems by means of analytical sociology

Agent-based modeling (ABM) has become a frequently used method to experiment with complex socio-technical systems, such as the transportation or the energy system, at the laboratory scale (Gramelsberger 2015; Van Dam et al. 2013). ABM has also been utilized as a tool to provide technology assessment with insights into alternative future pathways (Saam et al. 2019; Weyer and Roos 2017). An overview of various simulation frameworks and a comparison of different approaches to modeling transportation can be found in Weyer et al. (2022).

Some ABM frameworks such as the transportation simulation MATSim use simple decision algorithms, e.g., choosing the transport mode (car, bike, or public transport) that has performed best in the past, measured by time and costs (Horni et al. 2016). Again, this does not reflect the heterogeneity of real people's decisions, who may value time and costs differently.

Rooted in analytical sociology, the simulation framework SimCo (Simulation of the governance of complex systems) has been developed at TU Dortmund University since 2012 to investigate the effects of heterogeneous agents' interactions in infrastructure-based systems. It puts emphasis on the behavior of individuals and their modes of decision-making, which are shaped by subjective expectations (Kroneberg 2014). SimCo allows to explain why people behave differently, e.g., taking the car or the bike, and to simulate the interplay between individual everyday decisions at the micro level and system dynamics at the macro level.

The concept 'analytical sociology' has been coined by a group of researchers that try to model individual decision-making based on subjective preferences and (perceived) situational constraints (Esser 1993; Hedström and Swedberg 1996). Their goal is to understand the subjective reasoning of individuals and to explain chosen actions systematically as the result of a bounded-rational decision-making process, which can be reproduced by a mathematical algorithm. This algorithm includes situational factors ('Is there a bus station close to my home?') and preferences ('How much do I value traveling fast, cheap, environmentally friendly etc.?'), but also habits and routines ('Am I used to look for public transport opportunities, or do I automatically use my car?').

All three components of the decision-making algorithm are shaped by subjective perceptions, not only in the case of (subjective) preferences and (subjective) habits, but also in the case of (subjectively perceived) situational constraints ('Maybe, there is a bus station, but I have never noticed it.').

This decision algorithm, however, applies to all individuals similarly: Faced with various alternatives, e.g., taking the car or the bike, agents compare the utility (SEU) of each available option A (Figure 2), according to individual preferences (U) and probabilities (p) of achieving a goal by means of various action alternatives, and then typically choose the option that benefits them most (Konidari and Mavrakis 2007; Velasquez and Hester 2013).

$$SEU(A_i) = \sum_{j=1}^{n} p_{ij} \cdot U(O)_j$$

Fig. 2: SEU Calculation: U(O) = Utility of an expected result; p = probability of achieving a goal O. *Source: Konidari and Mavrakis 2007, p. 6247*

Table 1 illustrates this mathematical procedure referring to two fictitious people, who have different preferences (U): person A likes to travel fast, but does not care about costs, while person B is a money-saver, who is not interested in traveling fast. Probabilities (p) are the same for both persons here, but obviously could be adjusted due to different personal, residential or other situations, e.g., if fuel prices rise sharply, or cycle tracks are not available.

	Fast travel	Cheap travel	Utility
Person A			
Preferences (U)	10	4	
Car (p)	0.8	0.3	**9.2**
Bike (p)	0.3	0.8	6.2
Person B			
Preferences (U)	3	8	
Car (p)	0.8	0.3	4.8
Bike (p)	0.3	0.8	**7.3**

Tab. 1: Transport mode choice of two fictitious people (U values ranging from 1 to 10, p values from 0.0 to 1.0). *Source: authors' own compilation*

However, even this simple example shows that mode choice, based on mathematically calculated subjective utility, is very different: Person A takes the car (utility 9.2), while person B takes the bike (7.3). Parameters would have to be adjusted heavily to achieve behavioral change.

Hence, analytical sociology may contribute to making ABM more realistic, combining the idea of a general decision-making algorithm, that can be implemented into software, and the individuality of real people who decide autonomously based on subjective preferences, (subjectively) perceived situational constraints, and, finally, individual habits and routines that emerge as results of daily practices.

Based on the idea of individual decision-making, ABM allows to design complex systems by means of interactions of a large number of interdependent agents. Each agent in a transportation system changes the state of the system through their actions, e.g., choice of transport mode or route, and thus contributes to system dynamics. And each agent is vice versa influenced by the dynamic state of the system, e.g., traffic jam. This dynamic interplay of agents and system can hardly be investigated by theoretical reasoning. Experimental methods such as computer simulation provide researchers with the opportunity to observe self-organized processes of system dynamics and their emergent effects, which are hard to predict – and sometimes even surprising. Complex socio-technical systems often entail non-linear interactions, which can be investigated by running experiments on the computer and trying to interpret the results.

The simulation framework SimCo

The simulation framework SimCo has been developed to sustain and to push forward governance research, which mostly had been based on qualitative case studies. Focusing on governance and policy issues, SimCo does not pay much attention to physical dimensions of complex socio-technical systems such as the width of cycle tracks, but puts emphasis on the social dimensions of mobility or energy systems, i.e., the mobility or energy behavior of artificial agents who are designed to resemble real people's behavior. Therefore, the physical structure of the system is depicted as an abstract network, consisting of nodes and edges, which are freely parameterizable in their respective dimensions. As a general-purpose framework, SimCo allows conceptualizing edges as roads or cycle tracks and nodes as working places, residential places, or shopping malls.

According to analytical sociology, SimCo tries to explain system dynamics as the emergent result of the interactions of heterogeneous agents, making autonomous decisions. SimCo has been used for various experiments on risk management and system transformation, mostly in road transportation (Philipp and Adelt 2018; Weyer et al. 2019). Several what-if scenarios have been investigated analyzing the effects of external interventions on individual behavior, especially on mode and route choice (Adelt and Hoffmann 2017). Conducting experiments with simulation frameworks such as SimCo helps to better understand the real-time dynamics of complex socio-technical systems and to explain, e.g., why (various) people react (differently) to political measures intended to promote sustainable transformation (Weyer 2019).

By experimenting with various scenarios, e.g., of future mobility, experimenters can analyze the probability of success of different policies, such as banning cars with combustion engine

 https://doi.org/10.14512/tatup.32.1.56

(strong measure) or lowering prices for public transport (soft measure). ABM can be used as a method for assessing policy measures and their – sometimes unintended – effects, and also for predicting which policy strategies might have the biggest impact, e.g., in terms of sustainable transformation, and which would probably fail to achieve their goals.

Simulation of mobility in the Ruhr district

The Ruhr district with about five million inhabitants is one of the largest metropolitan areas in Germany, albeit with a rather atypical structure. Compared to Berlin, Hamburg, or Munich, there is no single center; instead the Ruhr district has a polycentric structure with a few big cities and several medium-sized or small towns. Like other metropolitan areas, the Ruhr district must cope with the challenge of sustainable transformation, especially in transportation, which is lagging behind other sectors.

InnaMoRuhr

In the InnaMoRuhr project (Concept of an integrated, sustainable mobility for the University Alliance Ruhr), three big universities collaborate in developing and implementing concepts for future transportation. The overall purpose of the project is to find out which policy measures or interventions may contribute to changing the mobility behavior of people studying and working at these three universities. One key element is the development of an agent-based model of the Ruhr district that can be used to test various scenarios, e.g., of promoting new modes of transport, such as bicycle traffic or car sharing, or even intermodal transportation, such as by bike to the station, by train to university.

As a first step, a survey has been conducted to collect data about typical mobility practices, but also about demands for future mobility: Participants were the students and employees of the three Ruhr universities. These data have been used to construct four scenarios, which were discussed and evaluated by members of all three universities at five scenario workshops. In parallel, the Ruhr model was technically implemented, partly relying on the NEMO model developed by other researchers, using the simulation framework MATSim to model and analyze urban transportation in the Ruhr district (Kaddoura et al. 2020; Ziemke et al. 2019). NEMO already depicts, e.g., the network of roads and tracks, the distribution of residential quarters, and the daily mobility behavior of people. However, since MATSim puts emphasis mostly on physical dimensions of transportation, such as travel time and costs, several sociological components had to be integrated, such as bounded-rational decision-making and, above all, the various agent types implemented in SimCo. The final step are real-world experiments to test those scenarios in practice that have proven promising in the scenario workshops as well as in simulation experiments.

Actor types

One major result of the survey (N = 10,782) was data about individual preferences (U values) and specific probabilities (p values) of achieving various goals with different modes of transport. As Table 2 shows, respondents were asked to indicate their travelling preferences (U values) in terms of dimensions like ‘fast’, ‘cheap’, or ‘safe’ by adjusting a slider in an online questionnaire (values ranging from 0 to 10). Numbers in column ‘mean’ show that respondents value speed (7.8) and reliability (8.1) highest and comfort (4.7) lowest.

Additionally, these data were used to create five distinct actor types by means of cluster analysis in the statistics program SPSS (Weyer 2022). Actors were clustered according to similarities in U values within-group and dissimilarities between-groups. As can be seen in Table 2, numbers differ remarkably between those five clusters. For example, actor type #4 (‘comfort-oriented’) rates comfort (+2.8) and safety (+1.4) much higher, and prices (–3.1) and environmental concerns (–2.1) much lower than the average (column ‘mean’). Conversely, actor type #5 (‘environmentally conscious and price sensitive’) rates prices

Dimension	Type 1	Type 2	Type 3	Type 4	Type 5	Mean
Fast	**–1.6**	**–0.8**	**1.0**	0.8	0.8	**7.8**
Cheap	**–0.6**	0.5	0.9	**–3.1**	**1.4**	6.3
Environmentally friendly	**1.7**	**0.9**	**–2.2**	**–2.1**	**2.1**	5.9
Comfortable	**–1.6**	0.8	0.3	**2.8**	**–2.2**	**4.7**
Safe	**1.7**	0.2	**–0.6**	**1.4**	**–2.7**	6.2
Reliable	0.5	**–1.9**	0.8	0.5	0.5	**8.1**
N =	2,081	2,522	2,808	1,598	1,747	10,782
Share	19.3 %	23.4 %	26.0 %	14.8 %	16.2 %	
Description of actor groups	(1) Risk averse and environmentally conscious; (2) Indifferent; (3) Pragmatic; (4) Comfort-oriented; (5) Environmentally conscious and price sensitive					

Tab. 2: Actor types, clustered by six dimensions of preferences (U values); high numbers (deviations from mean) are marked green, low numbers red.

Source: Weyer 2022, p. 20

https://doi.org/10.14512/tatup.32.1.56

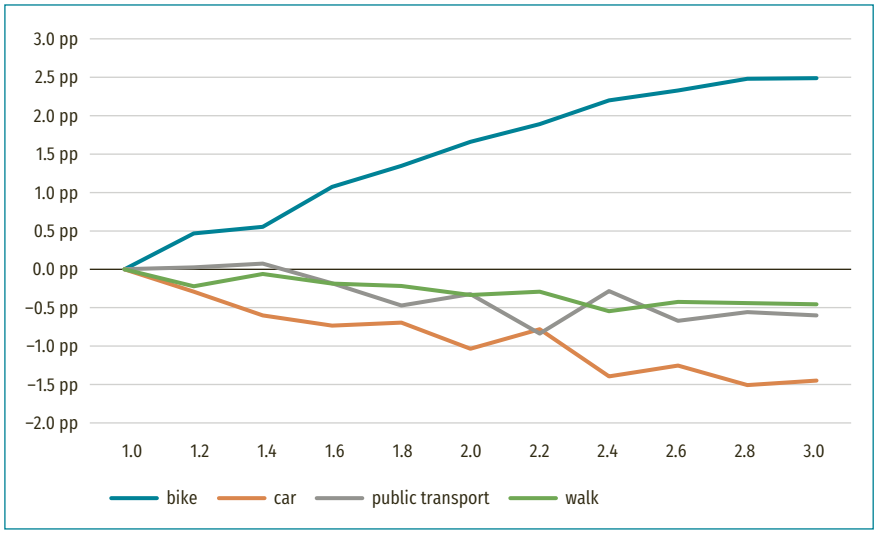


Fig. 3: Results of simulation experiments with raising bike comfort, depicting changes in mode choice compared to base scenario in percentage points (y-axis) related to depth of interventions (x-axis).
Source: Philipp et al. 2023

(+1.4) and environmental concerns (+2.1) much higher, but disregards comfort (–2.2) and safety (–2.7).

These actor types have been implemented as agent types in the MATSim-SimCo Ruhr model and experiments have been conducted to test three scenarios: (1) a mobility budget that allows people to gain experiences with alternative modes of transport; (2) a car sharing service that fills gaps in public transport between university and railway stations; (3) and a bicycle station that makes traveling by bike more convenient and safer. The following sections will present only the latter scenario.

Design of experiments

The basic NEMO model entails the whole Ruhr population, scaled down at one percent, resulting in 50,000 agents who behave according to the logic of MATSim, which is *not* based on analytical sociology. For our study, this global population of the Ruhr district was complemented by a university-specific population, representing twenty percent of about 130,000 members of three universities at Duisburg-Essen, Bochum and Dortmund. These 25,683 university agents were split up in the five agent types mentioned above, with additional variation in age, sex, profession, place of residence etc. They also behave according to the general logic of MATSim, but decision-making is based on subjective preferences and bounded rationality, referring to SimCo.

The original NEMO agents thus serve as a kind of 'background noise', affecting, e.g., the occupancy of roads, or public transport units used by both subsets of agents. However, the interventions within the scenarios only affect the university population – the main object of our study.

Experiments were conducted to test the willingness to change behavior, which is rather high, according to survey data: Faced with the option to take the bike to the station and the train to university (at a scale of 1 = no to 5 = yes), 73.7 percent of respondents were willing to make use of this option (sum of 4 and 5). However, the differences between agent types are remarkable: Comfort-oriented people rate this option much lower (3.32), and environmentally conscious and price sensitive people much higher (4.46) than the average (4.00).

To test the scenario 'bicycle station', comfort of cycling has been defined as the crucial parameter, the change of which might affect peoples' behavior, especially their willingness to take the bike for short or medium trips. Discussions in the scenario workshops revealed that many people like cycling but hesitate to use their own bike – especially expensive electric bikes – since safe storage at work or at the train station cannot be guaranteed.

Hence, experiments were conducted, increasing the parameter 'bike comfort' (p value, resulting from the survey) by using a factor from 1.0 (base value) to 3.0 (very high) in increments of 0.2. Since changing mobility patterns typically requires a combination of various policy measures, the costs of using cars for commuting were raised likewise (steps of 0.1) in a separate series of experiments, adjusting the parameter 'car costs' by using a factor from 1.0 (base value indicating low costs) to 0.0 (very expensive).

Finally, both measures have been combined, investigating the impact of simultaneously increasing bike comfort, e.g., by means of a bicycle station, and car costs, e.g. by introducing parking fees (Philipp et al. 2023).

Every single experiment depicts one typical day, starting in the morning and ending in the evening, including the daily mobility patterns of both the NEMO and the university population. According to MATSim's programming logic, the final daily plan is the result of 500 iterations, in which each agent adapts and optimizes its daily routine by trying various mobility options. This procedure has been executed with eleven parameter values per experiment (see x-axis of Figure 3).

Results of experiments

Figure 3 shows the results of simulation experiments with increased bike comfort using a factor from 1.0 (base value) to 3.0 (high comfort); changes in transport mode choice, indicated in percentage points (pp) compared to the baseline scenario are depicted (on the y-axis). Obviously, the intervention works, since transport mode choice shifts to bike (+2.5 percentage points at comfort level 3.0) at the expense of the other three modes. The reduction of car use (–1.4 pp) is highest compared to the baseline scenario.

https://doi.org/10.14512/tatup.32.1.56

In combination with a stepwise increase of (perceived) car costs, these numbers are even slightly higher: Bike usage rises by 3.6 pp, while car usage drops by 2.1 pp.

However, numbers differ when agent types are considered (Figure 4). Results are rather surprising, since it is not the two groups of environmentally concerned people (+1.9 and +1.1 pp for bike at intervention level 3.0 – filled columns), but mainly the groups of indifferent (+3.0 pp) and pragmatists (+4.1 pp) that contribute most to change by refraining from car use (–2.0 and –2.8 pp at level 3.0 – striped columns). At first glance, this seems to be counterintuitive, but if one considers that environmentally concerned people already are used to taking the bike, even if the weather is bad and storage is complicated or risky, then results are reasonable.

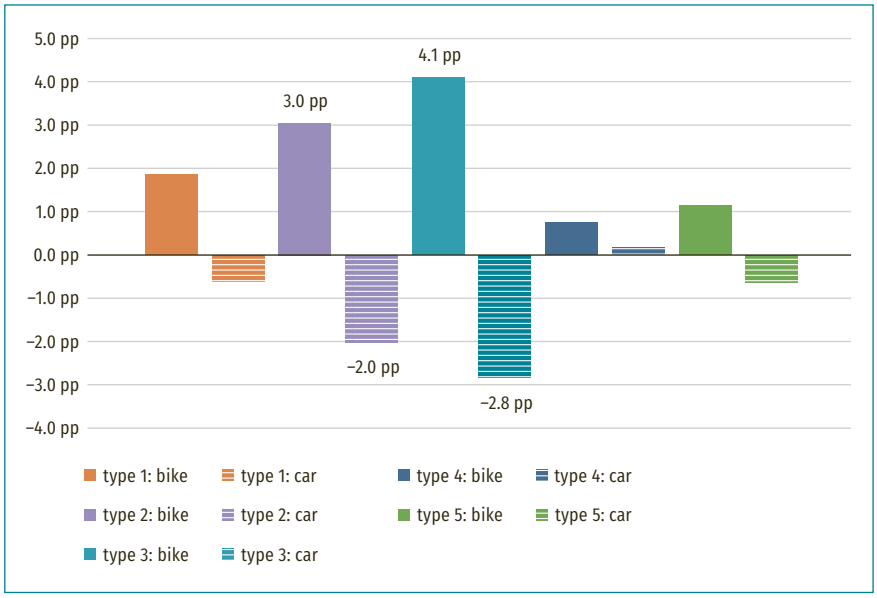


Fig. 4: Changes in transport mode choice compared to base scenario in percentage points (y-axis) of five agent types (x-axis) by raising bike comfort at level 3.0 (filled column: bike, striped column: car).
Source: Philipp et al. 2023

Hence, two agent types of environmentally concerned people (#1 and #5) reveal almost no change and only a minor trend of renouncing car use. This is partly because car ownership is much lower in those two groups: 63.6 and 58.5 percent, compared to an average of 71.7 percent and a maximum of 87.6 percent in the group of comfort-oriented (Weyer 2022, p. 23).

Agent type #4 of comfort-oriented people is hardly affected by increasing bike comfort, but agent types #2 (indifferent) and #3 (pragmatic) might be interesting target groups for policy measures fostering a transformation of the transportation system. Instead of addressing people who are already environmentally concerned and willing to travel climate friendly, politics should take a closer look at people that are indifferent or pragmatic but still – in contrast to comfort-oriented people – responsive to changing situational constraints, e.g., new opportunities (bicycle station) or restrictions (parking fees).

Numbers are slightly different, if two measures – raising bike comfort and car costs simultaneously – are combined. Bike use is even 2.0 percentage points higher among agent type #3 (pragmatic) at intervention level 3.0, at the expense of car use in this group (–1.7 pp). Again, comfort-oriented people (agent type #4) react differently: Car use goes up by 1.4 percentage points at intervention level 3.0, where cycling is convenient and car driving is extremely expensive – at first glance a counterintuitive effect as well.

This points at two findings: Many comfort-oriented people are willing to pay a high price to maintain their familiar habits. And second – still to be validated in more detail – temptation to use the car may rise, if streets are empty and congestions are rare, because many people have switched to bikes. This is a typical non-linear effect that can only be investigated by means of computer simulation.

Conclusion

Combining analytical sociology and agent-based models helps to better understand the variety of individual actions, the resulting system dynamics, and, finally, the different willingness of people to react to external interventions. Agent-based models, rooted in sociology, create a more realistic picture than other models, since they grasp everyday practices of people, who behave according to subjective preferences and make bounded-rational decisions.

Simulation experiments help to assess the impact of political interventions. Results of experiments with the MATSim-SimCo Ruhr model show that different agent types react very differently to measures intended to make transportation more climate friendly. Additionally, they help to identify those groups that may contribute most to sustainable transformation. Surprisingly, these are neither environmentally friendly actors, used to riding the bike, nor comfort-oriented actors, used to commuting by car, but two groups of rather indifferent or pragmatic people that are willing to change behavior.

Funding • Research in this article has been funded by the North Rhine-Westphalian Ministry of Transportation (OM 202018111).

Competing interests • The authors declare no competing interests.

References

Adelt, Fabian; Hoffmann, Sebastian (2017): Der Simulator „SimCo" als Tool der TA. Experimente zur Verkehrssteuerung. In: TATuP – Journal for Technology Assessment in Theory and Practice 26 (3), pp. 37–43. https://doi.org/10.14512/tatup.26.3.37

Dings, Christiane et al. (2022): Mathematische Modellierung des Verlaufs der SARS-CoV-2-Pandemie in den deutschen Bundesländern. Saarbrücken: Universität des Saarlandes. Available online at https://covid-simulator.com/wp-content/uploads/2022/08/Report_2022_08_17.pdf, last accessed on 05. 12. 2022.

Esser, Hartmut (1993): Soziologie. Allgemeine Grundlagen. Frankfurt a. M.: Campus.

Gramelsberger, Gabriele (2015): Computerexperimente. Zum Wandel der Wissenschaft im Zeitalter des Computers. Bielefeld: transcript.

Hedström, Peter; Swedberg, Richard (1996): Social mechanisms. In: Acta Sociologica 39 (3), pp. 281–308. https://doi.org/10.1177/000169939603900302

Horni, Andreas; Nagel, Kai; Axhausen, Kay (2016): The multi-agent transport simulation MATSim. London: Ubiquity Press.

Kaddoura, Ihab; Laudan, Janek; Ziemke, Dominik; Nagel, Kai (2020): Verkehrsmodellierung für das Ruhrgebiet. In: Heike Proff (ed.): Neue Dimensionen der Mobilität. Wiesbaden: Springer, pp. 361–386. https://doi.org/10.1007/978-3-658-29746-6_31

Konidari, Popi; Mavrakis, Dimitrios (2007): A multi-criteria evaluation method for climate change mitigation policy instruments. In: Energy Policy 35 (12), pp. 6235–6257. https://doi.org/10.1016/j.enpol.2007.07.007

Kroneberg, Clemens (2014): Frames, scripts, and variable rationality. An integrative theory of action. In: Gianluca Manzo (ed.): Analytical sociology. Norms, actions, and networks. Hoboken, NJ: Wiley, pp. 97–123. https://doi.org/10.1002/9781118762707.ch04

Philipp, Marlon; Adelt, Fabian (2018): Optionen der politischen Regulierung des Personenverkehrs. Ergebnisse einer Simulationsstudie. (Soziologisches Arbeitspapier 52/2018). Dortmund: TU Dortmund. http://dx.doi.org/10.17877/DE290R-18807

Philipp, Marlon; Adelt, Fabian; Weyer, Johannes (2023): Simulation des Mobilitätsverhaltens und möglicher Transformationspfade (Mobility Report 8/2023). Dortmund: TU Dortmund, in press.

Saam, Nicole; Resch, Michael; Kaminski, Andreas (2019): Simulieren und Entscheiden. Entscheidungsmodellierung, Modellierungsentscheidungen, Entscheidungsunterstützung. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-26042-2

Van Dam, Koen; Nikolic, Igor; Lukszo, Zofia (eds.) (2013): Agent-based modelling of socio-technical systems. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-4933-7

Velasquez, Mark; Hester, Patrick (2013): An analysis of multi-criteria decision making methods. In: International Journal of Operations Research 10 (2), pp. 56–66. Available online at http://www.orstw.org.tw/ijor/vol10no2/ijor_vol10_no2_p56_p66.pdf, last accessed on 10. 01. 2023.

Weyer, Johannes (2019): Die Echtzeitgesellschaft. Wie smarte Technik unser Leben steuert. Frankfurt a. M.: Campus.

Weyer, Johannes (2022): Mobilitätspraktiken und Mobilitätsbedarfe. Ergebnisse einer Befragung von Angehörigen der UA-Ruhr-Universitäten (Mobility Report 2/2022). Dortmund: TU Dortmund. Available online at https://innamo.ruhr/wp-content/uploads/2022/06/Report_02_Befragung_250422_final.pdf, last accessed on 05. 12. 2022.

Weyer, Johannes; Adelt, Fabian; Hoffmann, Sebastian (2019): Governance of transitions. A simulation experiment on urban transportation. In: Diane Payne et al. (eds.): Social simulation for a digital society. Applications and innovations in computational social science. Basel: Springer, pp. 111–120. https://doi.org/10.1007/978-3-030-30298-6_9

Weyer, Johannes; Adelt, Fabian; Philipp, Marlon (2022): Agent-based modelling of infrastructure systems. In: Jens Gurr, Rolf Parr and Dennis Hardt (eds.): Metropolitan research. Methods and approaches. Bielefeld: transcript, pp. 155–165. https://doi.org/10.14361/9783839463109-009

Weyer, Johannes; Roos, Michael (2017): Agentenbasierte Modellierung und Simulation. Instrument prospektiver Technikfolgenabschätzung. In: TATuP – Journal for Technology Assessment in Theory and Practice 26 (3), pp. 11–16. https://doi.org/10.14512/tatup.26.3.11

Ziemke, Dominik; Kaddoura, Ihab; Agarwal, Amit (2019): Entwicklung eines regionalen, agentenbasierten Verkehrssimulationsmodells zur Analyse von Mobilitätsszenarien für die Region Ruhr. In: Heike Proff (ed.): Mobilität in Zeiten der Veränderung. Wiesbaden: Springer, pp. 383–410. https://doi.org/10.1007/978-3-658-26107-8_29

PROF. DR. JOHANNES WEYER
has been Senior Professor for Sustainable Mobility at TU Dortmund University since 2022. Previously, he headed the Technology Studies Group for 20 years. Focus of work: Human-machine interaction, governance of complex systems, agent-based modeling and simulation. Publications: Vertrauen in digitale Technik, ZfS 2022; Die Echtzeitgesellschaft, Frankfurt a. M. 2019.

FABIAN ADELT
is a computer scientist and has been a research associate at TU Dortmund University (Technology Studies Group/Social Research Centre) since 2011. Focus of work: ABMS, Governance, socio-technical transitions in transportation and energy systems. Publications: Simulation of the governance of complex systems (SimCo), in: JASSS 2018; Modelling End-User Behavior and Behavioral Change in Smart Grids, in: Energies 2020.

MARLON PHILIPP
is an industrial engineer and joined TU Dortmund University in 2019 (Technology Studies Group/Social Research Centre). Focus of work: Sustainability and mobility research, modeling and simulation. Publications: Mikromobilität und Mobility-as-a-Service, in: Making Connected Mobility Work, 2021; Mobilitätspraktiken und Mobilitätsbedarfe in der UA Ruhr, in: Transforming Mobility – What Next? 2022.