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

Horvath, Lilla; Renz, Erich; Rohwer, Christian; Schury, Daniel

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Horvath, L., Renz, E., Rohwer, C., & Schury, D. (2023). Combining behavioral insights with artificial intelligence: New perspectives for technology assessment. *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis / Journal for Technology Assessment in Theory and Practice*, *32*(1), 43-48. https://doi.org/10.14512/tatup.32.1.43

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'.

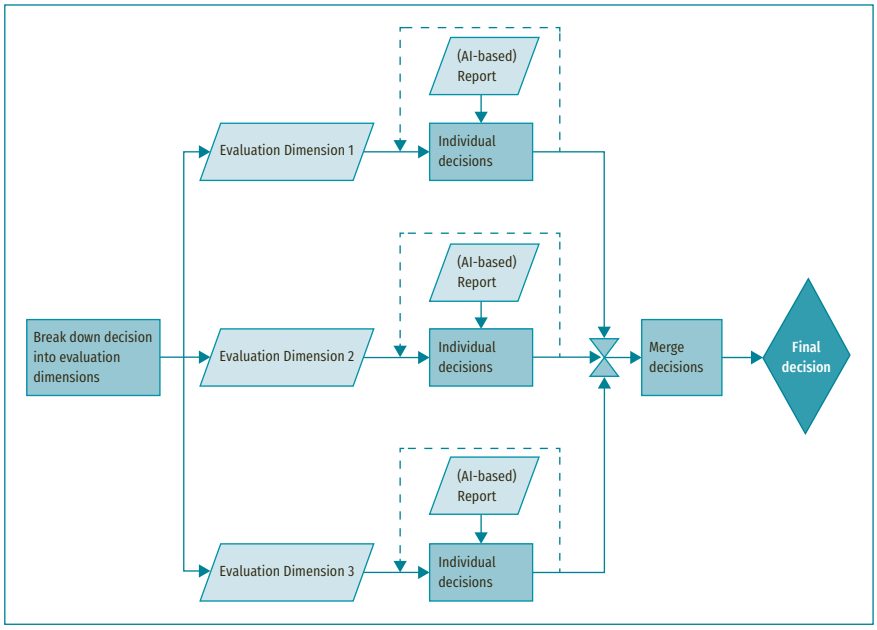


Fig. 1: Schematic flow chart of MAP. Rectangles represent actions; diamonds represent input. The dashed lines illustrate an (optional) repetition of the individual decisions phase. *Source: authors' own compilation*

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-

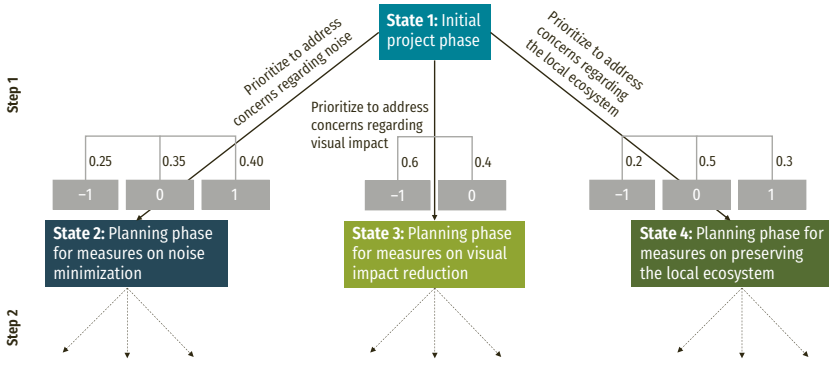
 https://doi.org/10.14512/tatup.32.1.43

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.