Sustainable Artificial Intelligence: In Search of Technological Resilience

Norbert Jastroch(✉)

MET Communications, 61352 Bad Homburg, Germany
norbert.jastroch@metcommunications.de

Abstract. Great demands are placed on the role of digital technology and artificial intelligence for sustainable development. External shocks, like the current pandemic, as well as creeping degradation, like the effects of carbon economy on the climate, need convincing concepts to control lasting negative effects on society, environment, and economy. This paper is intended to contribute to the search for ways of enabling resilience through technology.

High expectations are being put on data driven artificial intelligence in this respect. We consider that artificial intelligence tends to fall short of scientific rigor regarding cause-and-effect relations and discuss the inherent limitations of so-called formal systems that are at the bottom of artificial intelligence systems. We take into view what data analysis and reasoning can deliver regarding the discovery of empirical phenomena, arguing that targeted, reflective data reasoning can well help discover correlations worth further theoretical investigation. We suggest combining established methods of epistemic knowledge generation with data driven artificial intelligence, i.e. human intelligence with machine-based algorithmic intelligence, in support of advanced human-systems integration. For this concept of hybrid intelligence, we provide a procedural framework.

This methodological approach gets exemplified by the description of recently published cases of a technical application resp. Scientific practice, illustrating the potential of hybrid intelligence for the scientific as well as technical solution of problems. Concluding remarks finally draw the line to future work on sustainable artificial intelligence as a pathway to resilience delivered by technological means.

Keywords: Artificial intelligence · Formal systems · Data reasoning · Hybrid intelligence · Human-systems integration

1 Introduction

External shocks are often perceived as being sudden surprises. However, this only holds true under superficial observation or deficient reflection. So have the causes and effects of climate change been named since decades, but it took long to establish public insight, based upon scientific rigor. Likewise, the emergence of the SARS-CoV-2 virus and the subsequent global pandemic were only surprising as to the specific virus type – there have been numerous indications and warnings regarding potential virus pandemics since

Published by Springer Nature Switzerland AG 2023
F. Noël et al. (Eds.): PLM 2022, IFIP AICT 667, pp. 317–326, 2023.
https://doi.org/10.1007/978-3-031-25182-5_31

many years. And geopolitical tensions were virulent for long, including the risk of severe political and economic confrontation. While their extension into military action has been well observable, it was hardly believed until the day it turned into reality, leading to a major assault on a sovereign country which was a surprise to many in Europe. Although not coming out of the blue, these cases generated shock waves across the globe. They originated from technological (carbon economy), natural (virus evolution), or societal (geopolitical claims of imperial power) grounds.

And they are what the International Risk Governance Council (IRGC) calls emerging risks, resulting from the complexity of the concerned dynamic, non-linear systems they are part of. With focus on systemic risks in socio-economic systems, Helbing in [7] elaborated on the effects of complexity, in particular their cascading spread involved with network interactions. In their report [8] the IRGC built on his observations specifying several contributing factors that make up the fertile ground for risk emergence. Among these is the issue of technological advances. Changes in technology may become a source of risk if their impacts are not scientifically investigated in advance or surveyed after deployment - even more so if there is insufficient regulatory framework in place[1]. Thus, the IRGC are strongly arguing for ex ante as well as ex post risk assessment. The aim can be taken as a kind of sustainability evaluation, in the broad sense of securing sustaining, desirable implications whilst avoiding undesired side effects.

Interestingly, new technologies appear likewise to support the ability to adapt to future shocks. Brunnermeier [19], with a general societal perspective, points out that dealing with risk can either mean trying to avoid it, or to accept it in accordance with a framework of institutions, rules, and processes that are bound to enable recovery from external shocks. The first option remains constrained though, because total robustness, which covers any conceivable emergency, can hardly be realized as it would normally involve unacceptable high cost. The latter approach in fact finds increasing interest these days in the concept of resilience. A most relevant question then is to what extent can technology contribute to resilience, respectively become a driving force to it.

The currently most prominent technological field in this respect is digital transformation along with the resumed concept of artificial intelligence. The availability of massive data via digitization enables novel ways of empirical investigation. Techniques for their analysis do not only offer new approaches for applications in domains as diverse as health, mobility, manufacturing, agriculture, finance, energy, public administration etc. They also drive the development of what may be called Artificial Intelligence for Resilience, or Artificial Intelligence for Risk Governance.

This paper is intended to contribute to the search for technological ways of enabling resilience in that sense. The availability of massive data from digitization, along with powerful algorithms for analysis and reasoning, are the means to pro-actively assess risk scenarios and prepare for adequate responses of choice. We argue for the combination of human intelligence with algorithmic intelligence for the purpose of expanding theoretical knowledge from a theoretical as well as practical perspective. Our starting point

[1] For the technological field of Artificial Intelligence, and more general the digital transformation on its way these days, we have presented an investigation of risk mitigation matters in [10], comprising functional, societal, and cybersecurity risks. And their relation to regulative frameworks in the EU.

are reflections on formal systems, which are the methodological foundation of every scientific domain, and their limitation that must be accounted for in any application of artificial intelligence. A formal system is a set of axioms and rules for inference, and the resulting space of propositions, which altogether govern a domain of knowledge. It cannot be complete, free of contradiction, and closed at the same time. We then take into view what data reasoning, in particular machine learning can deliver for the discovery of empirical phenomena. As algorithmic machine-driven systems can become sources of novel security issues, need arises for precautionary searching of potential flaws by use of theoretical models. Respective procedural interrelations get illustrated in the framework for hybrid intelligence we suggest in the subsequent section, showing the combination of formal deductive methodology and probabilistic approaches. We continue by recurring to recently published cases of technical resp. Scientific examples showing the practical benefits, and close with concluding remarks as to future work on sustainable artificial intelligence.

2 Formal Systems are Limited

The scientific way of generating knowledge rests upon well-established processes that comprise the formulation of concepts, the generation of assumptions and their testing, the validation of findings and their alignment within a theory, and their incorporation into a coherent set of theorems and propositions which make up the body of domain knowledge [cf. 12]. Science is, in terms of methodology, extensively while not completely determined work, according to Kuhn [12]. Scientific progress is incremental and sometimes even disruptive when legacy paradigms get shifted – where the rationale of such a shift remains unclear in Kuhn's work [12]. Rovelli [16], on the contrary, argues that science and its progress work through continuity, not discontinuity. He identifies two origins of conceptual shifts in science: new data exerting decisive rationale for change, e. g. in the case of Kepler getting to his ellipses by mathematical analysis of empirical data of planet's courses, and informed investigation of contradictions within an existing theory, e. g. heliocentrism of our solar system. Following his observation of philosophy having contributed essentially to scientific development especially in the case of physics, Laplane et al. [13] more generally localize this impact of philosophy in the conceptual clarification and the critical assessment of assumptions or methods in a scientific discipline. We consider this a little deeper.

As mentioned, a scientific theory consists of a coherent set of theorems and propositions that are derived from a set of axioms with the help of rules of inference. The question of the validity of a theory in the sense of 'being true' has been subject to a wealth of debates in science theory. While the question 'What is truth?' could not be finally answered by philosophers throughout millennia up to now, Penrose [15] provided a comprehensive elaboration on the simplified question 'What is mathematical truth?'. The interest of our paper here are formal systems as part of artificial intelligence systems, i.e. abstract sets of axioms and rules for the purpose of inferring propositions to build a knowledge base to a certain domain. Mathematics is considered the perhaps most basic manifestation of such an axiomatic system, so it is safe to refer to Penrose's work [15].

Penrose [15] draws on the finding by Goedel formulated in his incompleteness theorem which applies to any formal system, consequently to any attempt of founding an artificial intelligence system on such a formalism. We content ourselves to refer the line of argument in brief without extensive detail: To ensure only valid propositions - in the sense of being mathematically proven - be derived, mathematical reasoning must be free of contradiction. Goedel showed that any such mathematical system, of whatever type, which is free of contradiction must include statements that are neither provable nor disprovable by the means allowed within the system. So full truth cannot be achieved within an axiomatic system by methods of proof. Penrose [15] pointed out that there is a way to get to the validity of a proposition he calls the reflection principle. By repeatedly reflecting upon the meaning, we can *see* it is true although we cannot derive it from the axioms. This *seeing* requires a mathematical *insight* that is not the result of deductive proof, or purely algorithmic operations which could be coded into some mathematical formal system. Admittedly, the status of this *insight,* as a mental procedure, remains unclear except its non-algorithmic nature. Its applicability appears as and insofar it leads to a coherent mathematical theory. At the very end, the consequence most interesting in the context of artificial intelligence - taken as a machine based algorithmic inference engine – may be: "…the decision as to the validity of an algorithm is not itself an algorithmic process" [15, p. 536]. Hence the inherent theoretical limitation of formal systems, which is depicted in Fig. 1: the object of human intelligence along with epistemics are domains of the real world, which include the physicalist part ('knowledgeable' in Fig. 1) and the non-physical qualities ('qualia')[2]. A formal system builds a proper part of a knowledgeable domain.

The algorithmic nature of purely machine-based Artificial Intelligence systems allows for formal procedures that cannot fully cover the related knowledge domain, not even the knowledgeable subset. Autonomous AI applications therefore will always have a blind spot area of propositions. The engineering of such AI systems needs to take care respectively, e.g. by ruling out their application in situations which might bear the risk of touching this area, or by calling in human guidance in such a situation. These issues get addressed under the concept of operational design domains, ODD, where functional constraints are introduced to avoid system states of that nature (cf. The case of autonomous vehicles engineering). Or by utilization of digital twins, which enable the investigation of such states for mitigation purposes (cf. The case of robotics).

[2] The difference between real domains and what is physically explainable is known as the ontological or epistemic gap – 'knowledgeable' then is a proper subset of reality. This philosophical distinction is not addressed in our context here.

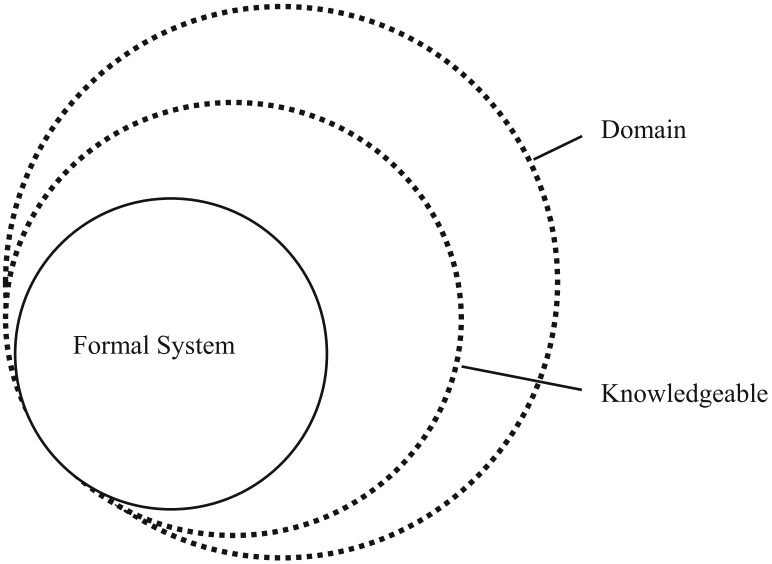

Fig. 1. The set of propositions of a formal system is a subset of the knowledgeable content of a domain which is a subset of the domain itself

3 Data Reasoning and Discovery

With the rise of digitization came data analytics as a source of innovation in science, industry, administration, and commercial marketplaces. Increasing utilization of massive data for automated decision making, machine learning and autonomous systems lead to several issues to be taken care of, among them data privacy (including data sovereignty, protection, safety) and data quality [cf. 9], and the validity of processes for their analysis and reasoning [cf. 10]. Uncontested, though, is the valuable potential of data driven artificial intelligence. The expectations are high: novel applications supporting societal and or economic progress; enhanced decision making based on algorithms; advancement of computational methods in sciences. In the context of this paper the latter both are of interest.

Our understanding of problems with our making decisions has been fundamentally enriched by the work of Kahneman et al. [11]. Deficient judgements appear to have two kinds of sources that are not related to the quality of information available: bias, the systematic deviation from neutral assessment, triggered by personal preferences; and noise, the statistical variance of judgements resulting from personal disposition[3]. They raise the question if and how bias and noise can be overcome with the help of metric criteria and parameter, and whether machine-based algorithms are per se more suited. Their answer is: they can be, it depends on the availability of parameters, their correct measurement, and the selection of parameters and measurement process being

[3] Eren and Mocan [6] provided an impressive empirical investigation of the correlation of unexpected football match losses with the length of sentences of judges, showing the impact of emotions in one domain on human decisions made in a completely unrelated domain.

free from bias and noise. However, Ludwig and Mullainathan [14] presented evidence of the fallibility of algorithms – be they controlled by humans or machines – depending on the algorithm as such, i.e. its fragility resulting from deficient construction. They resume that it is possible to reduce bias by well-built algorithms. If done right, artificial intelligence has the potential to undo human fallibility. They suggest human plus machine combination to be the best choice, though. A similar conclusion is drawn by Athey et al. [3]. Discussing when and how humans and machine-based algorithms should collaborate and who would be best to have formal decision authority, they argue for the AI system under optimal conditions as to data and algorithms employed, but for the combination of AI and human agent knowledge if that cannot be guaranteed.

The second option of harvesting the potential benefits of data reasoning lies in the generation of scientific novelty forced by data. It is not restricted to Physics if Rovelli [16] highlights sophisticated use of induction based upon accumulated empirical and theoretical knowledge as the most promising way forward in science. Regardless of whether massive data in a domain becomes available through digitization as such or by targeted experimental collection, it can be used to discover patterns and detect correlations that suggest new cause and effect relations (or propositions within the formal theory) to be tested for their validity. This is nothing new (recall the Kepler case mentioned in Sect. 2), but modern machine learning operations open a wealth of opportunities of that kind as they can find out conspicuous patterns in seconds instead of years of calculation. The basic process description is visualized in Fig. 2, as compared to Fig. 1 in Sect. 2. For exemplification, we present two recently published cases, one from pharmaceutical and another from historical research, in Sect. 5.

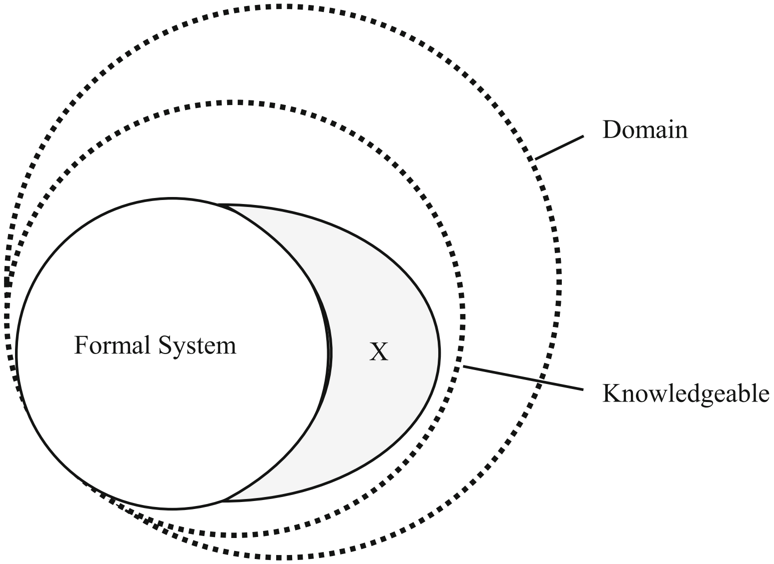

Fig. 2. A new empirical observation X stimulated the finding of a new proposition, what expands the formal system

4 Hybrid Intelligence and Its Potential

The limitation of Artificial Intelligence is a matter of principle, not of practicability [cf. 15]. Formal systems cannot completely cope with reality due to missing instantiations of e.g. ‘understanding’ and ‘meaning’, mental capabilities that are fundamentally human. But Artificial Intelligence enables the detection of correlations by data analysis techniques a human would not be capable of in terms of quantity of data to process. These techniques are at the heart of machine learning operations, with algorithmic procedures as their major building blocks. Algorithmic processes are the very kernel of Artificial Intelligence. Human intelligence, however, comprises these and also what Penrose called ‘reflection’ and ‘insight’. Both human and artificial intelligence are subject to bias, while artificial intelligence appears to be free of noise, other than humans. This is the rationale of combining human and artificial intelligence in the concept of hybrid intelligence (Fig. 3).

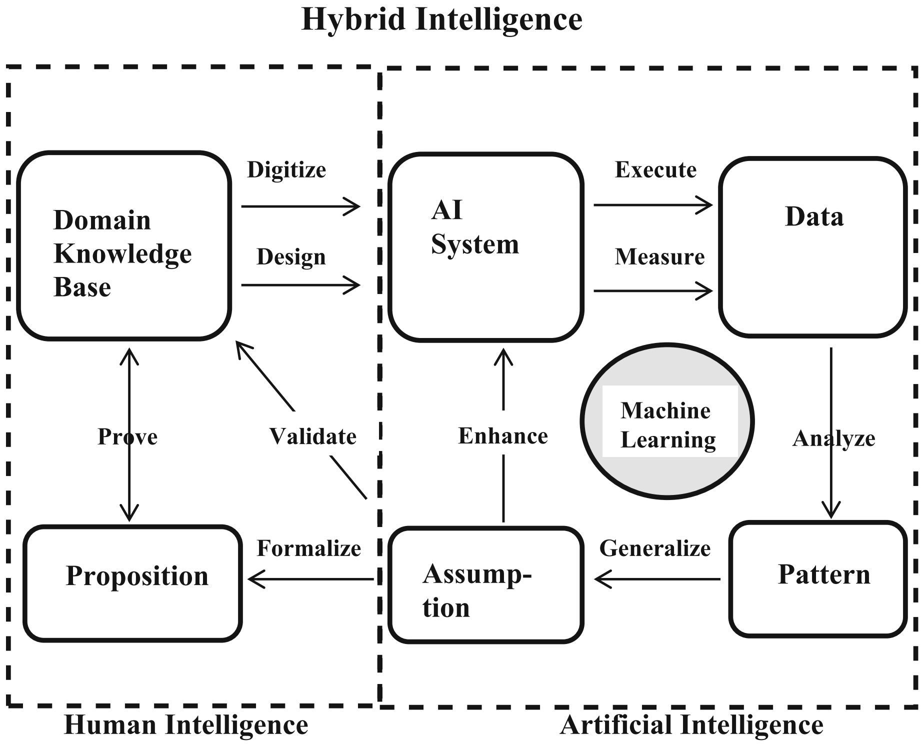

Fig. 3. Combining human and artificial intelligence

The Artificial Intelligence side of this concept is mainly building on data collection and analysis, and on pattern recognition and related reasoning. Altogether they make up the machine learning operation of the AI system. If such a system operates completely autonomous, the developing experience base can become instable and get out of order, as examples with text generation applications let show. Rudin [17] therefore strongly suggests machine learning to be tailored specifically to the domain of interest, thus enabling interpretability, as opposed to pure black box machine learning. This approach becomes focus in the concept of federated learning. Likewise, Athey [2] argues for big data applications in the policy field. She contests the usefulness of machine learning ‘off-the-shelf’, which is without understanding of underlying assumptions that require domain expertise to verify. Thus, she makes a strong point for supervised machine learning as it would be inherent to hybrid intelligence. And for computational social sciences Watts, Beck et al. [18] claim that data reasoning for predictive purposes must rest upon consistent estimation of causal relation, which requires the involvement of human domain knowledge.

Human supervision is represented in the left part in Fig. 3. It has controlling impact on the machine-based operations via digitization and design. Reversely, data reasoning on the AI side influences the advancement of the domain knowledge base by formalization and validation of assumptions drawn from patterns that were detected. It is worth noting here that in certain contexts of application human supervision can be important as it enables the purposeful introduction of bias which is explicitly desired. Cirillo et al. [5] discuss this for the field of precision medicine, where there is the need to consider sex related differences with physiological parameters or digital biomarkers when empirical patient data is collected for analysis.

5 Exemplifying Cases and Concluding Remarks

It is worthwhile to note that hybrid intelligence, as illustrated above, is not only meant to apply to scientific fields. It is appropriate in the industrial or public administration sphere, too, and also in commercial applications. In terms of innovation through knowledge generation however we expect the most valuable impact on science. Two cases of this kind that were subject to recent publications shall be described briefly here.

The first one is related to the COVID19 pandemic. When the SARS-CoV-2 virus emerged in the year 2019 and spread worldwide since 2020, a most unique innovation in pharmaceutics was accomplished in record time: the development of mRNA vaccines. They turned out the most powerful preventive measure against the COVID19 disease. However, numerous virus variants started to develop quickly and steadily. The mutant viruses showed significantly different levels of infective potential and severity of symptomatic illness. Soon questions arose about the effectiveness of vaccines and eventual need for adjustment. In the past the evaluation of new mutant risk rested on ex post observation of manifest infections, which required significant lead time for empirical investigation. Recently, though, as a novel approach a hybrid intelligence concept was put into action. The preprint of first results became available in December 2021. Beguir, Sahin et al. [4] describe their approach to the early computational detection of high-risk virus variants. It builds upon the combination of human domain expertise, as to the relation of specific virus structures and their potential implications, and the so called *in silico* assessment of their risk level with the help of supervised machine learning technique. Results of the machine learning application were reviewed by human domain experts as to their validity, and to enhance the model employed. The researchers report significant improvements regarding the time needed to detect dangerous variants, on average two months ahead of WHO sourced respective warnings. Furthermore, they demonstrated the applicability of their approach to real-time risk monitoring of mutations by an Early Warning System.

The second case is about historical research into the restoration of ancient text inscriptions that are damaged or remain preserved only partially, as most recently published [1]. The established scientific methods of epigraphy are constrained to the use of repositories of textual and contextual parallels and bring along high levels of generalization but low certainty of results. Assael et al. [1] applied a deep learning software based on neural networks that, with the help of human domain expertise, was carefully tailored to the epigraphic tasks to accomplish. They conducted experimental evaluation by applying their approach to a couple of ancient Greek inscriptions, using a specific metric to evaluate the performance achieved. They found substantial improvements of accuracy and speed of the restoration tasks under use of the combined human-machine intelligence concept. In fact, they report that the combination of human and artificial intelligence in an iterative process achieves significantly better results than human or artificial intelligence only.

These examples point to how technological resilience can be achieved with the help of hybrid intelligence approaches, be it pro-active risk governance, e. g. in health, or life-cycle advancement of ODD's, e.g. with autonomous vehicles, or enhancement of simulation models using digital twins, e. g. in smart manufacturing.

References

1. Assael, Y., et al.: Restoring and attributing ancient texts using deep neural networks. Nature **603**(7900), 280–283 (2022). https://doi.org/10.1038/s41586-022-04448-z
2. Athey, S.: Beyond prediction: using big data for policy problems. Science **355**, 483–485 (2017). https://doi.org/10.1126/science.aal4321
3. Athey, S., Bryan, K., Gans, J.: The allocation of decision authority to human and artificial intelligence. NBER working paper no. 26673 (2020). http://www.nber.org/papers/w26673
4. Beguir, K., Sahin, U., et al.: Early computational detection of potential high-risk SARS-CoV-2 Variants. bioRxiv preprint, 27 December 2021. https://doi.org/10.1101/2021.12.24.474095
5. Cirillo, D., et al.: Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digital Med. **3**(1) (2020). https://doi.org/10.1038/s41746-020-0288-5
6. Eren, O., Mocan, N.: Emotional judges and unlucky juveniles. Am. Econ. J. Appl. Econ. **10**(3), 171–205 (2018). https://doi.org/10.1257/app.20160390
7. Helbing, D.: Systemic risks in society and economics. International risk governance council workshop on emerging risks. Geneva, December 2009. (2010). https://irgc.org/wp-content/uploads/2018/09/Systemic_Risks_Helbing2.pdf
8. International Risk Governance Council: The Emergence of Risks: Contributing Factors. Geneva (2010)
9. Jastroch, N.: Trusted artificial intelligence: on the use of private data. In: Nyffenegger, F., Ríos, J., Rivest, L., Bouras, A. (eds.) PLM 2020. IAICT, vol. 594, pp. 659–670. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62807-9_52
10. Jastroch, N.: Applied artificial intelligence: risk mitigation matters. In: Junior, O.C., Noël, F., Rivest, L., Bouras, A. (eds.) Product Lifecycle Management. Green and Blue Technologies to Support Smart and Sustainable Organizations: 18th IFIP WG 5.1 International Conference,

PLM 2021, Curitiba, Brazil, 11–14 July 2021, Revised Selected Papers, Part I, pp. 279–292. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-94335-6_20

11. Kahneman, D., Sibony, O., Sunstein, C.: Noise. Siedler Verlag, München (2021)
12. Kuhn, T.S.: Die struktur wissenschaftlicher revolutionen. 23rd impression. Suhrkamp Verlag, Frankfurt a. M (2012)
13. Laplane, L., et al.: Why science needs philosophy. In: PNAS 5 March 2019, vol. 116, no. 10, pp. 3948–3952 (2019). https://doi.org/10.1073/pnas.1900357116
14. Ludwig, J., Mullainathan, S.: Fragile algorithms and fallible decision-makers: lessons from the justice system. J. Econ. Perspect. **35**(4), 71–96 (2021). https://doi.org/10.1257/jep.35.4.71
15. Penrose, R.: The Emperor's New Mind. Revised impression 9, Oxford Landmark Science. Oxford University Press (2016)
16. Rovelli, C.: Physics needs philosophy. philosophy needs physics. Found. Phys. **48**(5), 481–491 (2018). https://doi.org/10.1007/s10701-018-0167-y
17. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Preprint. arXiv: 1811.10154v3. 22 Sep 2019 (2019)
18. Watts D.J., Beck, E., et al.: Explanation, Prediction, and Causality: Three Sides of the Same Coin? OSF Preprint. 31 Oct 2018. https://doi.org/10.31219/osf.io/u6vz5
19. Brunnermeier, M.K.: Die resiliente Gesellschaft. Aufbau Verlage, Berlin (2021)