Künstliche Intelligenz

Herausgegeben von
dem Bundesministerium für Umwelt,
Naturschutz, nukleare Sicherheit und
Verbraucherschutz
und FRAUKE ROSTALSKI

Mohr Siebeck

Künstliche Intelligenz

Künstliche Intelligenz

Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?

herausgegeben von

dem Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

und

Frauke Rostalski

Mohr Siebeck

Frauke Rostalski, geboren 1985; Studium der Rechtswissenschaften in Marburg; 2011 Promotion Rechtswissenschaften; 2017 Promotion Philosophie; seit August 2018 Inhaberin des Lehrstuhls für Strafrecht, Strafprozessrecht, Rechtsphilosophie und Rechtsvergleichung an der Universität zu Köln.
orcid.org/0000-0002-5606-3639

Veröffentlicht mit Unterstützung des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz.

ISBN 978-3-16-161298-5 / eISBN 978-3-16-161299-2
DOI 10.1628/978-3-16-161299-2

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über *http://dnb.dnb.de* abrufbar.

Das Buch wurde von Laupp & Göbel in Gomaringen gesetzt, auf alterungsbeständiges Werkdruckpapier gedruckt und von der Buchbinderei Spinner in Ottersweier gebunden.

Printed in Germany.

Inhaltsverzeichnis

What is "Good" Training Data?

Data Quality Dimensions that Matter for Machine Learning

Felix Neutatz / Ziawasch Abedjan

I. Introduction

Machine learning (ML) is becoming prevalent in almost all areas of our everyday life and enables artificial intelligence (AI) technologies that affect humans significantly with applications in personalized medicine,[1] automated credit rating,[2] and justice systems,[3] to name a few. As a result, it is vital to make sure that decisions and results that originate from ML are of high quality and explainable. One of the imminent problems in current AI systems is that they amplify societal bias, which harms minorities and other protected groups disproportionately. At this point, it is necessary to keep in mind that humans are not only at the receiving end of ML systems but also influence these systems at various major stages of their development and production life cycle. At each of these stages, there is a potential to spill over societal bias into the ML model. First, the data that is used to train these systems can originate from humans. For instance, Microsoft presented a chatbot that continuously learned from the interaction with its users. Some adversaries fed the bot with offensive content that in turn changed the behavior of the chatbot to be racist and sexist.[4] Second, humans develop and configure the embedded ML production pipelines. Technical choices of the developer can introduce technical bias into the entire process from data preparation to model creation.[5] Third, humans analyze and interpret the data and the ML model predictions. After the model returns the predictions, humans still have to decide how to design the thresholds

---

[1] *Emmert-Streib, F./Dehmer, M.*, A machine learning perspective on personalized medicine: An automized, comprehensive knowledge base with ontology for pattern recognition, Mach. Learn. Knowl. Extr., vol. 1, no. 1, pp. 149–156, 2019.

[2] *Bono, T./Croxson, K./Giles, A.*, Algorithmic fairness in credit scoring, Oxford Review of Economic Policy, vol. 37, no. 3, pp. 585–617, 2021.

[3] *Mayson, S. G.*, Bias in, bias out, Yale LJ, vol. 128, p. 2218, 2018.

[4] *Lee, P.*, Learning from Tays introduction, 2016 [online]. Available: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

[5] *Schelter, S./He, Y./Khilnani, J./Stoyanovich, J.*, FairPrep: promoting data to a first-class citizen in studies on fairness-enhancing interventions, EDBT, 2020, pp. 395–398; *Donini, M./Oneto, L./Ben-David, S./Shawe-Taylor, J./Pontil, M.*, Empirical risk minimization under fairness constraints, NeurIPS, 2018, pp. 2796–2806.

for different decisions and how to rate different types of mispredictions.[6] False negatives and false positives have different implications and societal cost. For example, detecting a cancer diagnosis by mistake (false positive) has a different impact than missing a cancer case (false negative). Humans have to understand these trade-offs and configure the model and the resulting decision support system, accordingly. One of the main pillars of responsible data management is to ensure good training data.[7] It is widely understood that "good" training data yields high ML model accuracy. Traditionally, good training data has the following characteristics: correct, complete, up-to-date. The data needs to be correct because if annotations are incorrect ML models learn incorrect patterns. Likewise, missing properties inside a dataset reduce the overall expressiveness of a model. Finally, data has to be up-to-date because if we train a model on data from 10 years ago and apply it to data of today, temporal concept shifts change the distributions that cause mispredictions. The larger the amount of such data, the easier it is for the model to generalize and differentiate noise from actual trends. With the maturity of ML algorithms and systems and their application on real-world use cases, good data also has to satisfy additional characteristics: it has to be representative of different groups of a population and free from historic bias. A representative sample of all data is important to ensure high model accuracy for all population groups. For example, in many image datasets, people with dark skin are underrepresented. This misrepresentation has led to poor prediction performance for this group. E. g. Twitter was focusing white people's faces while cropping the faces of people with dark skin[8] or Google[9] and Facebook's[10] models predicted people with black skin as gorillas. While representation requires explicit modeling of different population groups, for some use cases such information leads to amplification of discrimination based on historic biases towards certain groups. One example of such a case is a system for supporting hiring decisions. It turned out that the majority of the historic data contained male hires. Therefore, the model learned that gender was an accurate signal for the prediction task and its predictions discriminated against women.[11] This example shows that it is crucial to

---

[6] *Stoyanovich, J. / Howe, B. / Jagadish, H. V.*, Responsible data management, PVLDB, vol. 13, no. 12, pp. 3474 – 3488, 2020.

[7] *Stoyanovich / Howe / Jagadish* (Fn. 6).

[8] *Chowdhury, R.*, Sharing learnings about our image cropping algorithm, 2021 [online]. Available: https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

[9] *Vincent, J.*, Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech, 2018 [online]. Available: https://www.theverge.com/2018/1/12/16882408/google-racistgorillas-photo-recognition-algorithm-ai.

[10] *Mac, R.*, Facebook apologizes after a.i. puts 'primates' label on video of black men, 2021 [online]. Available: https://www.nytimes.com/2021/09/03/technology/facebook-ai-raceprimates.htm.

[11] *Dastin, J.*, Amazon scraps secret ai recruiting tool that showed bias against women, 2018 [online]. Available: https://www.reuters.com/article/usamazon-com-jobs-automation-insight-idUSKCN1MK08G.

consider concepts, such as equality of outcome through demographic parity, to protect minority groups. Often the problem is linked to so-called sensitive attributes that are historically correlated but causally irrelevant for the outcome of a prediction task, such as gender, race, or religion. A naive solution would be to simply drop those sensitive attributes. This way the algorithm is blind across these dimensions. Unfortunately, this approach does not work well in practice because the model might learn the affiliation of persons to minority groups from other attributes – so-called proxy attributes. For instance, *Selbst* showed that zip code is a proxy attribute for race in the US.[12] In this paper, we focus on the aspect of fair data engineering. We can approach fairness in two ways: from the individual perspective or the group perspective. Individual fairness ensures equal treatment for people with similar characteristics. Group fairness ensures equal treatment across groups – no group is disadvantaged. At the first glimpse, individual and group fairness might contradict each other in some cases. For example, a higher qualified person from a majority group misses an opportunity, which fell to the top candidate from a minority group despite lower qualification scores. However, Reuben makes the case that with careful consideration, we can avoid such dilemmas.[13] For instance, in the case of financial lending, we start by asking what is the purpose of the algorithm – here, which persons should get a credit? The only important point is that a person can pay back the credit in the end. The second question that has to be answered is what kind of assumptions do we have for the data. E. g. where does the data about creditworthiness come from and is this data meaningful or is it prejudicial? The next question is which characteristics should be used? To estimate creditworthiness, one can rely on income, securities, and assets. Finally, we need to understand whether there is historic or structural discrimination in the data. For example, maybe a structurally poor place of residence is associated with credit denial because despite having high potential to be creditworthy they lack the historic evidence that equals to candidates from more prosperous places. After answering all these questions, one can understand the domain-specific bias problem and can address it with a technical solution. The examples show that the original goal of good data to maximize model accuracy is not sufficient. Instead, one has to consider the problem as a multi-objective problem or more practically a maximization problem under constraints. Constraints are minimum thresholds on metrics that capture novel quality dimensions of ML. These quality dimensions prominently include fairness, privacy (e. g. GDPR compliance), or explainability. In this article, we will briefly discuss the implication of data quality with regard to traditional dimensions as well as novel population-level dimensions on AI and surface the current state-of-the-art technologies that reduce bias in ML technol-

[12] *Selbst, A. D.*, Disparate impact in big data policing, Georgia law review, vol. 52, p. 3373, 2017.

[13] *Binns, R.*, On the apparent conflict between individual and group fairness, FAccT, 2020, pp. 514–524.

ogies and their training data. We demonstrate one particular technology, which is designed to deal with sensitive attributes and their proxies and conclude with a set of suggestions for reconciling socio-technical aspects of algorithmic fairness.

II. How Data Quality impacts AI

ML relies on – and is "programmed" by – training data.[14] Thus, the quality of the training data is fundamental toward robust and accurate models, and ultimately toward useful and reliable ML-based applications.[15] Thus, there is no surprise that data and ML engineers report spending a tremendous amount of time on preparing datasets for ML applications.[16] Traditional dimensions of data quality include completeness, correctness, and freshness of datasets and are per se independent of the downstream ML application. Completeness of a dataset captures to which degree a dataset is populated with content. Incomplete datasets typically miss attribute values either because of negligence in the data creation phase or because certain data values are not known. As ML-based systems are required to understand associations between given properties and the target property, it is easy to see that missing values might reduce the performance. Similarly, it is well understood that incorrect values in the data might negatively impact the accuracy of an ML system. Research on data quality has so far led to a large number of methods, heuristics, and systems that support data quality improvement through data cleaning, which comprises the identification and correction of data quality problems, such as identifying missing values and imputing them. While most of the existing work focuses on cleaning independent of the application – here ML – there are novel sparks towards ML-dependent cleaning techniques. Traditional techniques typically rely on error models to detect duplicate or outlying values, external constraint information (e. g., business or integrity constraints), or human assessment and input (e. g., to recommend repairs or cleaning examples). The separation of data cleaning and the application can lead to several problems. First, it is hard for users to antic-

---

[14] *Neutatz, F. / Chen, B. / Abedjan, Z. / Wu, E.*, From cleaning before ml to cleaning for ml, IEEE Bulletin, 2021.

[15] *Lee* (Fn. 4); *Li, P. / Rao, X. / Blase, J. / Zhang, Y. / Chu, X. / Zhang, C.*, CleanML: a study for evaluating the impact of data cleaning on ml classification tasks, ICDE, 2021; *Baylor, D. / Breck, E. / Cheng, H.-T. / Fiedel, N. / Foo, C. Y. / Haque, Z. / Haykal, S. / Ispir, M. / Jain, V. / Koc, L.* et al., Tfx: A tensorflow-based production-scale machine learning platform, SIGKDD, 2017, pp. 1387–1395.

[16] *Deng, D. / Fernandez, R. C. / Abedjan, Z. / Wang, S. / Stonebraker, M. / Elmagarmid, A. K. / Ilyas, I. F. / Madden, S. / Ouzzani, M. / Tang, N.*, The data civilizer system, CIDR, 2017; *Agrawal, A. / Chatterjee, R. / Curino, C. / Floratou, A. / Godwal, N. / Interlandi, M. / Jindal, A. / Karanasos, K. / Krishnan, S. / Kroth, B. / Leeka, J. / Park, K. / Patel, H. / Poppe, O. / Psallidas, F. / Ramakrishnan, R. / Roy, A. / Saur, K. / Sen, R. / Weimer, M. / Wright, T. / Zhu, Y.*, Cloudy with high chance of DBMS: a 10-year prediction for enterprise-grade ML, CIDR, 2020; *Kandel, S. / Paepcke, A. / Hellerstein, J. M. / Heer, J.*, Enterprise data analysis and visualization: An interview study, TVCG, vol. 18, pp. 2917–2926, 2012.

ipate which cleaning routines matter for the downstream ML routine, which will unequivocally lead to a waste of resources and user time. Interestingly, improving a dataset could in fact degrade the outcome of the downstream ML model.[17] For instance, it is not clear whether a partial improvement of the data quality might lead to other inconsistencies with systematic errors that had been exploited via downstream neural networks. In a prior study, we manually curated clean and dirty versions of the FAA Flights delay[18] and U.S. Census datasets[19]. Each dataset served a different prediction task. We showed that whether or not cleaning is beneficial heavily depends on the application. For the Flights dataset, cleaned training data improves the model accuracy on both clean and dirty test data. However, cleaning the Census training data actually degrades the model accuracy on dirty data. In fact, training and testing on dirty data is as accurate as training and testing on clean data, yet requires no effort. This experiment provides evidence that cleaning is not a local "one-and-done" process. In fact, the appropriate cleaning intervention is dependent on the type of error as well as the rest of the application, and should be approached from this perspective. Consequently, all of the complexities inherent in modern ML applications become complexities that affect how data is cleaned. In prior work, we argued that data cleaning needs to take an end-to-end application-driven approach that integrates cleaning throughout the ML application.[20] In contrast to traditional data cleaning, there are approaches that are embedded within the ML development phase, which focuses on model development, training, and evaluation. Although the ML community has developed a multitude of robust model designs and training techniques,[21] it is often better to directly address errors and biases in the data.[22] Methods that are applied in this phase leverage the ML models and validation data to identify both the data points for cleaning and the appropriate cleaning routine that directly improve the ML accuracy. So far the discussed methods and problems related to objective and factual problems with the data. However, data quality in the context of AI spans other dimensions as well. In particular, traditional means of cleaning do not count in population-level problems. A recent study found out that existing data imputation methods rely on the maximum likelihood of values skewing the result of imputation towards majority groups and amplifying misrepresentation of minority groups in the data.[23] In fact, the data quality dimension of fairness with regard to metrics on demographic par-

---

[17] *Amershi, S./Begel, A./Bird, C./DeLine, R./Gall, H. C./Kamar, E./Nagappan, N./Nushi, B./Zimmermann, T.*, Software engineering for machine learning: a case study, ICSE, 2019, pp. 291–300.

[18] *Mahdavi, M./Abedjan, Z.*, Baran: Effective error correction via a unified context representation and transfer learning, PVLDB, vol. 13, no. 11, pp. 1948–1961, 2020.

[19] *Li/Rao/Blase* et al. (Fn. 15).

[20] *Neutatz/Chen/Abedjan/Wu* (Fn. 14).

[21] *Li, J. Z.*, Principled approaches to robust machine learning and beyond, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, USA, 2018.

[22] *Li/Rao/Blase* et al. (Fn. 15).

[23] *Schelter/He/Khilnani/Stoyanovich* (Fn. 5).

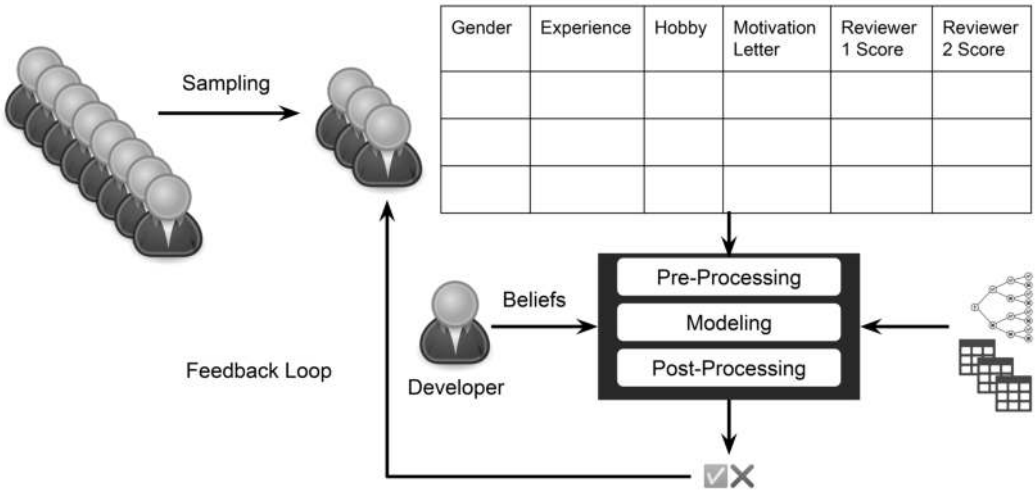

Figure 1: Fairness in the ML Workflow.

ity has gained more attention in recent years. In the next section, we discuss how existing work tries to assess and enforce fairness in ML applications.

III. How Data leads to Discrimination

Machine learning finds patterns in the data to make predictions for new unseen data. If this data is biased, the extracted patterns are likely to be biased, too. These biased patterns lead to predictions that discriminate. Biases sneak into an ML application across the entire ML workflow, which consists of data collection, pre-processing, modeling, and post-processing. *Friedman* and *Nissenbaum* identified three main types of bias: pre-existing, technical, and emergent.[24] To describe these types of bias in more detail, we explain them with the help of the workflow for an ML hiring application as shown in Figure 1. Pre-existing bias has its roots in the current beliefs of society and is generally independent of the ML application. For instance, the fraction of women in the German parliament in 2021 is only 35 %,[25] or the well-known phenomenon – the gender pay gap – that women earn significantly less than men.[26] So, even if we would be able to gather data on all people in the world, an accurate representation of existing circumstances in our society would let our ML application to reproduce discrimination. For instance, if a model leverages the salary to predict whether a person can get a credit or not,

[24] *Friedman, B./Nissenbaum, H.*, Bias in computer systems, ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347, 1996.

[25] Wikipedia, Frauenanteil im Deutschen Bundestag seit 1949, 2021 [online]. Available: https://de.wikipedia.org/wiki/Frauenanteil im Deutschen Bundestag seit 1949.

[26] *Blau, F. D./Kahn, L. M.*, Understanding international differences in the gender pay gap, Journal of Labor economics, vol. 21, no. 1, pp. 106–144, 2003.

the gender pay gap makes the salary a proxy attribute for gender and therefore, the model will discriminate against women. Consider our running example of an ML-driven hiring application. To address pre-existing bias, the data scientists have to first assess the types of risks with regard to discrimination and how existing circumstances influence the outcome. For instance, one could follow the concept of substantive equality of opportunity that only compares people to other people with the same circumstances.[27] Next, the data scientist has to define the characteristics that should be allowed to judge the applicants. Assume, the available data is the gender, the experience, the hobby, the motivation letter, and two reviewers' scores. The data scientist would carefully remove attributes that are not supposed to influence the outcome. The application should be blind with regard to attributes, such as gender. Ideally, the scientist would also identify proxies of such attributes that strongly correlate with their values. For example, the data scientists discover that the hobby is a proxy attribute for gender. So, they remove this attribute. Furthermore, they realize that reviewer 2 prefers applicants from certain universities. Therefore, they group the score by the university and normalize it. This way, they make sure that they can use the reviews without introducing bias. Finally, the data scientists have to make sure that, for all specified features, the necessary data is available across all groups because missing values might introduce new bias. This concept is known as feature equity.[28] After formulating the beliefs and identifying the attributes, one might think that, now, we can develop any algorithm and do not need to worry about fairness anymore because the data does not contain any sensitive or proxy attributes. This blindness approach is quite common. However, in many cases, we still have to measure the bias based on the formulated beliefs because the underlying data or the following algorithm might introduce bias. Technical bias is introduced or enforced by the developed ML application. Technical bias can occur at any stage of the ML workflow, ranging from sampling, preprocessing, modeling, and post-processing. In the sampling phase, the data scientists have to choose from a large number of previous applicants who they choose as a foundation to learn who to hire. An example of bias in sampling is an Amazon ML application, which "learned" that female candidates are less likely to succeed at the company because their data contained mainly male candidates.[29] So, we have to ensure to draw a representative sample across all demographic groups. This ensurance is known as representation equity.[30] So, the data scientists might access

---

[27] *Zehlike, M./Yang, K./Stoyanovich, J.*, Fairness in ranking: A survey, CoRR, vol. abs/2103.14000, 2021.

[28] *Jagadish, H. V./Stoyanovich, J./Howe, B.*, The many facets of data equity, Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference, Nicosia, Cyprus, March 23, 2021, ser. CEUR Workshop Proceedings, C. Costa and E. Pitoura, Eds., vol. 2841. CEUR-WS.org, 2021 [online]. Available: http://ceur-ws.org/Vol-2841/PIE+Q 6.pdf.

[29] *Dastin* (Fn. 11).

[30] *Jagadish/Stoyanovich/Howe* (Fn. 28).

data about previous applicants in the company and whether they perform well or poorly later on. Needless to say that it is critical that the HR department ensures that the pool of applicants is as diverse as possible and that the interview process is as fair as possible. Otherwise, the first step of sampling is likely to introduce bias already. Pre-processing is another step that might introduce bias. It transforms the data into a machine-understandable format. These transformations include data cleaning and augmentation. For instance, *Schelter* et al. showed that missing value imputation such as mean value imputation can enforce bias, especially because minority groups are more likely to avoid disclosing sensitive information.[31] Therefore, developers have to keep in mind that any data transformation that joins or filters the data might change the data distribution and introduce or amplify bias. An example of how data augmentation can introduce bias can be explained by the transformation that translates the textual motivation letter into numeric form. For instance, one can leverage a neural network that models the language based on the data extracted from Wikipedia to transform the textual motivation letter into a numerical vector. However, 87 % of the contributors of Wikipedia are men and therefore consciously or unconsciously introduce bias.[32] Therefore, data scientists have to check all components and libraries that they include into the workflow. In the post-processing phase, data scientists modify the thresholds when to hire a candidate or not. Depending on the company, the policy allows for more false positives or false negatives. False negatives are applicants that are not hired but would have been great employees if they would have had a chance to prove themselves. False positives are applicants that are hired but turn out to not fit the company. Modifying these thresholds, the data scientists always keep track of fairness because different thresholds might affect minorities in different ways. After some time with the company, the hired applicant's data can be used to train a new model. However, the data scientists keep monitoring to avoid any reinforcing biases in this feedback loop. For instance, they should be careful to avoid survivorship bias. In the Second World War, the US army analyzed where planes were shot to strengthen the parts that are more likely to be in danger. Abraham Wald realized that it was best to strengthen the parts that in the analysis were unscathed because all the planes that they analyzed actually made it back to safety.[33] So, in our case, we run into the danger of only including data points about applicants that were actually hired. While one can remove preexisting bias and technical bias in a system during implementation, emergent bias arises only in a context of use, e. g. by incorrect use or distribution shifts over time. For instance, after finishing a hiring application for the German branch of a company, a manager wants to use

---

[31] *Schelter/He/Khilnani/Stoyanovich* (Fn. 5).

[32] *Torres, N.*, Why do so few women edit wikipedia, Harvard Business Review, vol. 2, 2016.

[33] *Wald, A.*, A method of estimating plane vulnerability based on damage of survivors, Center for Naval Analyses, 1980.

the same application internationally. However, the school systems differ significantly and might discriminate against some nationalities. Therefore, it is important that whenever an application is used in a new context, one has to analyze all potential biases again. This concept is also known as output equity.[34] We showed that bias can be introduced at multiple stages of any ML application. As far as the ML and data science community is aware of these problems there have been attempts and technical solutions for several of the aforementioned pitfalls.

IV. State of the Art

Algorithmic bias reduction has been approached from different perspectives. In this section, we briefly survey existing approaches for fair ML and discuss fundamentals of feature engineering, which will be important to present our take on bias reduction in ML. Algorithmic bias reduction can be categorized threefold: by where in the ML pipeline they address the issue, how they measure bias, and what types of bias they address. Algorithmic bias reduction can be implemented at three stages: during preprocessing, in-processing, or post-processing. Pre-processing approaches[35] reduce the bias by modifying the data, e. g. shifting the probability distributions in the data,[36] selecting features,[37] or weighting features[38]. In-processing approaches[39] reduce the bias in the model, e. g. by modifying

---

[34] *Jagadish/Stoyanovich/Howe* (Fn. 28).

[35] *du Pin Calmon, F./Wei, D./Vinzamuri, B./Ramamurthy, K. N./Varshney, K. R.*, Optimized pre-processing for discrimination prevention, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 3992–4001; *Feldman, M.* et al., Certifying and removing disparate impact, Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), 2015, pp. 259–268; *Zhang, L./Wu, Y./Wu, X.*, A causal framework for discovering and removing direct and indirect discrimination, in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 3929–3935; *Asudeh, A./Jagadish, H. V./Stoyanovich, J./Das, G.*, Designing fair ranking schemes, in Proceedings of the International Conference on Management of Data (SIGMOD), 2019, p. 1259–1276.

[36] *du Pin Calmon/Wei/Vinzamuri/Ramamurthy/Varshney* (Fn. 35); *Feldman* et al. (Fn. 35); *Zhang/Wu/Wu* (Fn. 35).

[37] *Galhotra, S./Shanmugam, K./Sattigeri, P./Varshney, K. R.*, Fair data integration, CoRR, vol. abs/2006.06053, 2020.

[38] *Asudeh/Jagadish/Stoyanovich/Das* (Fn. 35).

[39] *Schelter/He/Khilnani/Stoyanovich* (Fn. 5); *Kilbertus, N./Rojas-Carulla, M./Parascandolo, G./Hardt, M./Janzing, D./Scholkopf, B.*, Avoiding discrimination through causal reasoning, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 656–666; *Nabi, R./Shpitser, I.*, Fair inference on outcomes, Proceedings of the Conference on Artificial Intelligence (AAAI), 2018, pp. 1931–1940; *Russell, C./Kusner, M. J./Loftus, J. R./Silva, R.*, When worlds collide: Integrating different counterfactual assumptions in fairness, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 6414–6423; *Perrone, V./Donini, M./Kenthapadi, K./Archambeau, C.*, Fair bayesian optimization, CoRR, vol. abs/2006.05109, 2020.

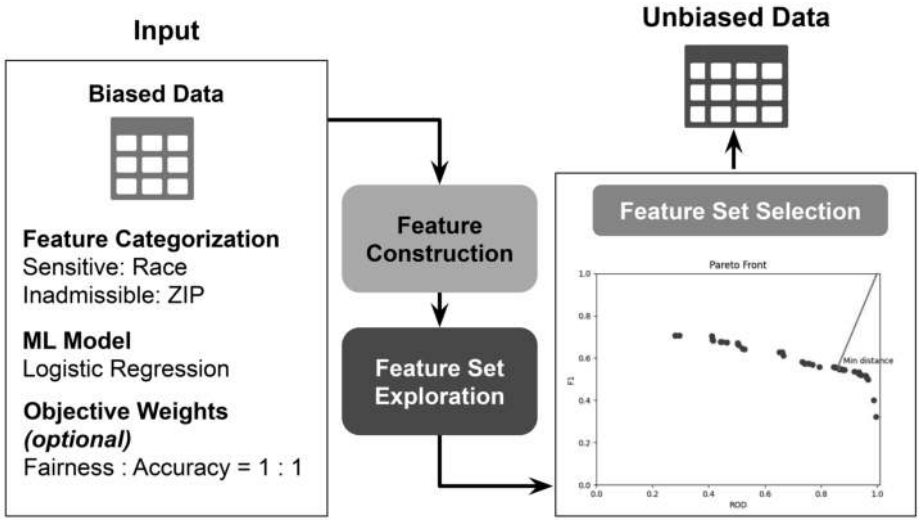

Figure 2: Architecture of the Fairness Explorer.

the model's loss.[40] Another example is to optimize the model's hyperparameters, which describe the configurations of a model for a setting, based on a fairness metric.[41] Post-processing approaches[42] reduce the bias in the final model predictions by applying transformations.[43] The second way to differentiate bias reduction approaches is to compare their fairness metrics. The two main approaches to measuring fairness are associational and causal. Associational approaches measure fairness in the model's predictions between groups of the sensitive feature. Three representative fairness criteria that are used by such approaches are independence, separation, and sufficiency.[44] Independence describes the concept of returning the same prediction for two similar individuals that only differ with respect to their sensitive attribute, such as religion, race, or gender. However, this metric ignores potential correlations between the group and the prediction target. Separation allows the score and the sensitive attribute to correlate to the extent that is justified

[40] *Kamishima, T./Akaho, S./Asoh, H./Sakuma, J.*, Fairness-aware classifier with prejudice remover regularizer, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), ser. Lecture Notes in Computer Science, vol. 7524, 2012, pp. 35–50.

[41] *Perrone/Donini/Kenthapadi/Archambeau* (Fn. 39).

[42] *Hardt, M./Price, E./Srebro, N.*, Equality of opportunity in supervised learning, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323; *Woodworth, B. E./Gunasekar, S./Ohannessian, M. I./Srebro, N.*, Learning non-discriminatory predictors, Proceedings of the Conference on Learning Theory (COLT), vol. 65, 2017, pp. 1920–1953.

[43] *Hardt/Price/Srebro* (Fn. 42); *Corbett-Davies, S./Pierson, E./Feller, A./Goel, S./Huq, A.*, Algorithmic decision making and the cost of fairness, Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), 2017, pp. 797–806.

[44] *Barocas, S./Hardt, M./Narayanan, A.*, Fairness and Machine Learning, 2019, http://www.fairmlbook.org.

by the target variable. So, both the true positive rate and the false positive rate have to be the same across groups. Sufficiency formalizes that the score subsumes the sensitive attribute to predict the target. This criterion is achieved by calibration.[45] Causal approaches model the causal relationships between the attributes in the data and measure in this way the influence of the sensitive attribute on the model predictions. Causality-based approaches[46] avoid paradoxical conclusions and provide more principled reasoning about the influence of the sensitive features on the model's predictions.[47] The third way to differentiate bias reduction is to compare which types of bias the approaches address. We differentiate between pre-existing bias, technical bias, and emergent bias [23].[48] Pre-processing approaches can only address preexisting bias because they do consider upstream transformations or algorithms. In-processing and post-processing approaches can address technical bias additionally because they both address the bias at the last component of the technical workflow – the model. To the best of our knowledge, there have been no technical solutions proposed that address emergent bias.

V. Fairness Explorer

To optimize an ML application for accuracy as well as a fairness metric often leads to a trade-off between the two optimization criteria. One of the goals of research in algorithmic fairness is to reduce the necessity of trade-offs by keeping both success criteria high. To this end, we developed the Fairness Explorer,[49] an approach to reduce bias in the downstream prediction task through pre-processing. In short, our approach excludes biased signals of sensitive attributes and their proxies by generating non-critical signals that can substitute them. The naive approach of removing data bias would be to only remove all sensitive attributes and all proxy attributes. This approach would unbias the data. However, this way, we also remove signals and correlations that reduce model accuracy significantly. Therefore, our idea was to extract as much unbiased information as possible from sensitive and proxy attributes without introducing new bias. Specifically, we apply transformations on the original attributes to obtain new attributes. For instance, in our example in Section 3, Reviewer 2 prefers applicants from specific universities. As explained before, we can unbias the score by using a transformation that

---

[45] *Barocas/Hardt/Narayanan* (Fn. 44).

[46] *du Pin Calmon/Wei/Vinzamuri/Ramamurthy/Varshney* (Fn. 35); *Salimi, B./Rodriguez, L./Howe, B./Suciu, D.*, Interventional fairness: Causal database repair for algorithmic fairness, SIGMOD, 2019, pp. 793–810.

[47] *Salimi/Rodriguez/Howe/Suciu* (Fn. 46).

[48] *Friedman/Niessenbaum* (Fn. 24).

[49] *Salazar, R./Neutatz, F./Abedjan, Z.*, Automated feature engineering for algorithmic fairness, Proc. VLDB Endow., vol. 14, no. 9, pp. 1694–1702, 2021.

normalizes the score based on overall university rankings and scale scores correspondingly. This way, the bias in the score is removed without forcing us to entirely remove the reviewer's assessment. Based on all these newly generated attributes, the Fairness Explorer picks a set of these attributes by optimizing for both fairness and accuracy. As illustrated in Figure 2, the Fairness Explorer has three main components: feature construction, feature set exploration, and feature set selection. First, the user specifies the dataset and the sensitive attributes in a causal framework, which identifies proxies to the sensitive attributes. The feature construction component creates new features from the original features by recursively applying aggregation and scaling transformations. The feature set exploration component implements a two-phase exploration strategy of various constructed feature sets. As smaller and less complex feature sets generalize better, the exploration starts to search for the minimal set of features that maximizes accuracy. In the second phase, starting with the feature set that maximizes accuracy, we remove one feature at a time to find a feature set that maximizes fairness. Then, the feature set selection component computes the feature sets that describe the best trade-offs between fairness and accuracy. The user can then pick the feature set that displays a trade-off best fitting to the use case at hand. Finally, the Fairness Explorer returns the preprocessed, bias-reduced dataset to the user.

VI. Conclusion

In this paper, we briefly discussed the role of data in AI. We pointed out that effective AI systems heavily rely on large sets of high-quality training data whereas quality encompasses dimensions that assess the quality of individual data points as well as dimensions to assess the fitness for use of the entire dataset. One of the latter dimensions refers to fairness in data. We laid out how the impediment of fairness, namely bias, is subtly introduced into data, be it as a distorted reflection of reality or an accurate reflection of a reality that entails unattended discrimination. In particular, we pointed out that to get rid of bias one has to be aware of pre-existing, technical, and emergent bias. We showed that researchers have proposed a series of best practices, metrics, and algorithms to measure, monitor, and avert bias in the data and ultimately the ML application. We showcased the Fairness Explorer as one example of such systems. In summary, we believe that reliable and safe AI systems require developers to know about all attack vectors of bias – such as pre-existing bias, technical bias, and emergent bias. We concur with the previously expressed view[50] that there is no one-size-fits-all technical solution to fairness in data. Each ML application requires a domain-specific but holistic

[50] *Stoyanovich/Howe/Jagadish* (Fn. 6).

approach to address bias based on the beliefs and values of the society. To make fair ML applications the norm, we have to step up as a society and work together. It is more important than ever to involve people of different demographic groups in the ML development process. The chances to identify hidden bias in our applications will increase and we will be in a position to implement countermeasures.

Einsatz von KI im Versicherungssektor

Rechtliche und ethische Herausforderungen

Christian Armbrüster

I. Einführung

Der Einsatz von Künstlicher Intelligenz (KI) im Bereich der Privatversicherung bietet für alle Beteiligten in verschiedener Hinsicht Vorteile. Genannt sei hier nur beispielhaft die KI-gestützte Betrugserkennung, welche keineswegs allein den Versicherern zugutekommt, sondern über niedrigere Prämien sowie höhere Überschussbeteiligungen und Beitragsrückerstattungen auch der Gesamtheit der redlichen Versicherungsnehmer.[1] Zugleich sind allerdings mit dem Einsatz KI-gestützter Algorithmen aus Verbrauchersicht auch einige teils beachtliche Risiken verbunden. Dies gilt insbesondere für die Antragsablehnung aufgrund einer KI-basierten Risikoprüfung sowie für die Schadensregulierung mittels KI, sofern problematische Kriterien herangezogen werden.

Der Begriff der KI ist nicht fest umrissen. Die „High Level Group on Artificial Intelligence" der Europäischen Kommission hat folgende Definition vorgeschlagen:

„Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. […]"[2]

Weitere Definitionen setzen etwas andere Akzente.[3] Nach dem gegenwärtigen Stand der digitalen Technologie geht es, wenn von KI die Rede ist, meist um das sog. maschinelle Lernen, in fortgeschrittenem Stadium als Deep Learning bezeichnet.[4] Hierbei erhält die Software eine bestimmte Aufgabe und versucht

---

[1] Soweit im Folgenden aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet wird, werden damit alle Geschlechter erfasst.

[2] https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines. Alle Online-Quellen wurden zuletzt am 13.09.2022 abgerufen.

[3] S. dazu etwa *Klar*, BB 2019, 2243 f.

[4] Näher *Niederée/Nejdl*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Rechtshandbuch Künstliche Intelligenz und Robotik, 2020, § 2 Rn. 24 ff., 56 ff.

diese zu lösen. Dies erfordert eine aufwändige Trainingsphase: Zunächst bedarf es einer bestimmten Eingabeinformation, beispielsweise in Gestalt eines Bildes. Bei der Analyse dieser Information wird sodann eine schrittweise Fehlerreduzierung betrieben. Ziel ist die Annäherung an sog. Schwellwerte (thresholds, verstanden als kleinster Wert einer Größe, der als Ursache einer erkennbaren Veränderung ausreicht), ab denen Neuronen Ausgabeinformation erzeugen. In der Lernfähigkeit des KI-gestützten Algorithmus wird zu Recht das zentrale Unterscheidungsmerkmal zu schlichten Computerprogrammen erblickt.[5]

Ein bekanntes Beispiel im Bereich der Mustererkennung[6] bietet eine Sequenz von Fotos, welche die Übergänge von einem Blaubeer-Muffin zum Gesicht eines Chihuahuas zeigen.[7] Dabei hat sich die Fehlerquote im Laufe der letzten Jahre kontinuierlich verringert. Im Versicherungssektor liegt ein praktisches Anwendungsfeld beispielsweise in der KI-gestützten Schadensanalyse. Dabei wird das Foto eines Kfz-Blechschadens in der Ausgabeinformation als neuer Schaden oder aber als bereits vor dem Verkehrsunfall vorhandener früherer Schaden identifiziert. Auf dieses Beispiel wird im Zusammenhang mit der Kfz-Schaden-App noch zurückzukommen sein (s. sub III).

Beim sog. Supervised Learning sind die Eingabedaten bereits klassifiziert. Beispielsweise enthält die Eingabeinformation die Tatsache, dass es sich um ein bestimmtes Kfz handelt. Einen Schritt weiter geht das sog. Unsupervised Learning. Hierbei klassifiziert das KI-Modell die eingegebenen Daten selbst. Nach der Trainingsphase folgt die Einsatz- und Monitoring-Phase, in deren Verlauf der Algorithmus zugleich weiter trainiert werden kann. Im Bereich der sog. Tiefenanalyse werden unstrukturierte Daten wie Tonaufnahmen, Bilder oder Videos mittels KI in strukturierte (Tabellen-)Daten umgewandelt. Dies ermöglicht eine automatisierte Analyse, bei der mehr Details erfasst werden als von den menschlichen Sinnen.

Darüber hinaus wird KI häufig auch mit der sog. Robotik in Verbindung gebracht, also der Ergänzung motorischer Intelligenzleistungen des Menschen. Dieses Einsatzfeld von KI ist insbesondere im Bereich von Medizin und Pflege, aber auch bei der industriellen Produktion bedeutsam. Für die hier interessierenden Fragen von Chancen und Risiken des KI-Einsatzes im Versicherungssektor kann die Robotik hingegen außer Betracht gelassen werden.

Von entscheidender Bedeutung für die Effektivität des Einsatzes von KI sind generell zwei Faktoren, nämlich Datenquantität und Datenqualität. Was die *Datenquantität* angeht, so gilt: Je mehr Daten verfügbar und eingespeist sind, desto fehlerfreier sind die Schwellwerte. Dabei kommt mithin das sog. Gesetz der großen

---

[5] S. dazu etwa *Rüfner*, in: Dederer/Shin (Hrsg.), Künstliche Intelligenz und juristische Herausforderungen, 2021, 15 (18).

[6] Näher *Stiemerling*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kapitel 2.1 Rn. 7 ff.

[7] https://blog.cloudsight.ai/chihuahua-or-muffin-1bdf02ec1680.

Zahl zum Tragen, welches im Versicherungssektor auch in anderer Hinsicht, nämlich für den Risikoausgleich im Kollektiv, von entscheidender Bedeutung ist.[8] Zur *Datenqualität* wiederum ist folgendes festzuhalten: Je exakter, vollständiger und aussagekräftiger die Daten sind, desto fehlerloser sind die Schwellwerte. Insgesamt lässt sich damit sagen: Je mehr und je bessere Daten vorliegen, umso komplexere Fälle kann KI sachgerecht bearbeiten.

Im Folgenden sollen zunächst die verbraucherrelevanten Einsatzfelder von KI im Versicherungssektor vorgestellt werden (sub II), bevor zunächst auf die rechtlichen Anforderungen (sub III), den Nutzen (sub IV) und sodann auf die Risiken (sub V) eingegangen wird, die damit für Verbraucher verbunden sind. Diese Risiken verdienen im Hinblick auf einen effektiven Verbraucherschutz und den entsprechenden Regulierungsbedarf (sub VI), aber auch auf die Erhöhung der Akzeptanz von KI-basierten Entscheidungen bei Verbrauchern, besondere Aufmerksamkeit. Abschließend wird ein Fazit und Ausblick geboten (sub VII).

II. Einsatzfelder

1. Vertragsanbahnung

Die möglichen Einsatzfelder von KI in der Privatversicherung sind vielfältig.[9] Schon im Vorfeld eines Vertragsschlusses, wenn es um die Ermittlung des Versicherungsbedarfs und die darauf aufbauende vorvertragliche bedarfsgerechte Beratung (§§ 6 Abs. 1 S. 1, 61 Abs. 1 S. 1 VVG) geht, kann KI zum Einsatz kommen. Liegt dem Versicherer im nächsten Schritt ein Vertragsantrag des Versicherungsnehmers vor, so tritt er in die Phase der Risikoprüfung und -bewertung ein. Hierbei spielen die vom Antragsteller in Erfüllung der vorvertraglichen Anzeigepflicht gem. § 19 Abs. 1 VVG gemachten Angaben zu risikorelevanten Umständen eine Rolle.

Bereits die Ausarbeitung der Antragsfragen, welche der Versicherer dem Antragsteller unterbreitet, kann KI-gestützt sein. Zu beantworten sind nämlich allein solche Fragen, die sich auf gefahrerhebliche Umstände beziehen.[10] Welche Umstände dies sind, „die für den Entschluss des Versicherers, den Vertrag mit dem vereinbarten Inhalt zu schließen, erheblich sind" (§ 19 Abs. 1 S. 1 VVG), kann auf unternehmensinternen oder aber auf externen Datenanalysen beruhen. Als Beispiel für Letzteres sei das Zonierungssystem für Überschwemmungsrisiko und Einschätzung von Umweltrisiken (ZÜRS Geo)[11] genannt. In engem Zusammen-

---

[8] S. dazu *Armbrüster*, Privatversicherungsrecht, 2. Aufl. 2019, Rn. 267.

[9] Beispiele bei *Spieleder/Robers*, ZfV 2019, 538 ff.; s. auch *Armbrüster/Prill*, ZfV 2020, 110 ff.

[10] Prölss/Martin/*Armbrüster*, VVG, 31. Aufl. 2021, § 19 Rn. 2 ff.

[11] https://www.gdv.de/de/themen/news/-zuers-geo----zonierungssystem-fuer-ueberschwemmungsrisiko-und-einschaetzung-von-umweltrisiken-11656.

hang mit der Risikoanalyse und den dafür maßgeblichen Faktoren steht – als ein weiterer unternehmensinterner Vorgang im Vorfeld des Vertragsschlusses – die Prämienkalkulation des Versicherers. Auch für sie kann der Einsatz von KI eine bedeutsame Rolle spielen.

2. Produktgestaltung

Auch bei der Gestaltung von Versicherungsprodukten kann KI zum Einsatz kommen. Ein Beispiel bieten die sog. Pay-as-you-drive-Tarife (Telematiktarife) in der Kfz-Versicherung, bei denen aufgrund von Datenauswertungen zum Fahrverhalten eine risikoadäquate Prämie bemessen wird.[12]

3. Schadensregulierung

Haben sich die bisher genannten Einsatzfelder von KI auf das Stadium vor Eintritt eines Versicherungsfalls bezogen, so ist ein weiterer praktisch bedeutsamer Bereich die Regulierung von Schadensfällen durch den Versicherer. Auch hier gibt es in der Praxis bereits jetzt wichtige Anwendungsfelder von KI. Dabei geht es typischerweise um Versicherungsfälle, die massenhaft auftreten und bei denen ein allseitiges Interesse an zügiger Regulierung besteht.

Ein Beispiel bietet die Abwicklung von Kfz-Unfällen, bei denen allein die Regulierung von Bagatellschäden, insbesondere von Blechschäden, in Rede steht.[13] Hier eröffnen einige Versicherer dem Versicherungsnehmer in der Kfz-Versicherung die Möglichkeit, Fotos des Schadens aufzunehmen und diese über eine Kfz-Schaden-App an den Versicherer weiterzuleiten. Es folgt eine KI-gestützte Analyse der Schadensmeldung, bei der auf Datenbanken mit Schadenshergängen, Fotos und Reparaturinformationen zurückgegriffen wird. Diese Analyse bezweckt es, die Abgrenzung zu Vorschäden sowie eine Einschätzung des Schadensbehebungsaufwands vorzunehmen. Ergibt sich dabei, dass ein regulierungsfähiger Schaden in einer bestimmten Größenordnung vorliegt, erfolgt ein automatisiertes schnelles Regulierungsangebot mit einer konkret benannten Entschädigungssumme, deren rasche Auszahlung in Aussicht gestellt wird. Rechtlich handelt es sich dabei um ein Angebot des Versicherers zum Abschluss eines Regulierungsvertrags. Nimmt der Versicherungsnehmer das Angebot an, so erhält er zügig die Auszahlung, und die Regulierung ist damit abgeschlossen.

---

[12] Näher *Ph. Koch*, VersR 2020. 1413 ff.; zu datenschutzrechtlichen Fragen (auch) in diesem Kontext *Haupt*, NZV 2020, 501 ff.

[13] https://versicherungsmonitor.de/2019/11/15/blechschaeden-mit-ki-regulieren/.

4. Dunkelverarbeitung

Schon bei der Vertragsanbahnung, aber auch im weiteren Verlauf eines bereits begründeten Versicherungsverhältnisses, spielt die KI-gestützte sog. Dunkelverarbeitung unproblematischer Informationen eine wichtige Rolle. Dies betrifft im Vertragsanbahnungsstadium etwa stark standardisierte Anträge im Massengeschäft, im laufenden Vertragsverhältnis Adressänderungen oder die Beantwortung von Standardfragen und im Regulierungsstadium einfachere Schadensberechnungen.

5. Betrugserkennung

Ein weiteres bedeutsames Einsatzfeld KI-gestützter Analysen ist die Betrugserkennung. Nach wie vor verursacht Versicherungsbetrug für die Versicherungswirtschaft hohe Schadensaufwendungen. Mithilfe von KI kann es gelingen, den Tätern rascher und häufiger auf die Spur zu kommen.

6. Vertrieb von Versicherungsprodukten

Beim Vertrieb von Versicherungsprodukten durch Versicherer, selbstständige Versicherungsvertreter sowie Makler kann der Einsatz von KI dazu beitragen, die sog. Churn Rate zu reduzieren. Darunter versteht man den Anteil derjenigen Versicherungsnehmer, welche ihren Vertrag kündigen. Mittels KI-gestützter Datenanalysen können aus Sicht des Vertriebs etwa bereits im Vorfeld einer möglichen Kündigung dem Versicherungsnehmer attraktive Angebote gemacht werden, um die Bereitschaft, die Vertragsbindung aufrechtzuerhalten, zu stärken. Aus Sicht des Verbraucherschutzes wichtige Grenzen werden diesem Einsatzfeld freilich durch die Anforderungen des Datenschutzrechts gesetzt. Dasselbe gilt für den Einsatz von KI zum sog. Up- und Cross-Selling, also für die Nutzung der vorhandenen Kundendaten, um höherpreisige oder andersartige Versicherungsprodukte zu vertreiben (z. B. einen umfangreicheren als den bereits bestehenden Gebäudeversicherungsschutz oder eine Kapitallebensversicherung zusätzlich zur Kfz-Haftpflichtversicherung).

Der Einsatz von KI im Vertrieb kann mithin aus Verbrauchersicht dazu dienen, bedarfsgerechten Versicherungsschutz zu erlangen oder fortzuführen und dadurch Deckungslücken in der privaten Risikovorsorge zu schließen.[14] Zugleich sind dann, wenn sich das Kollektiv der Versicherungsnehmer vergrößert, auch für den einzelnen Versicherungsnehmer Kostenvorteile zu erwarten. Freilich ist gerade beim Einsatz von KI im Vertrieb zu beachten, dass letzterer sich am Kundeninteresse auszurichten hat. Dies kommt in dem in §§ 6 Abs. 1 S. 1, 61 Abs. 1 S. 1

[14] Zur Plansicherungsfunktion der Privatversicherung s. *Armbrüster*, Privatversicherungsrecht, 2. Aufl. 2019, Rn. 219 ff.

VVG enthaltenen Gebot bedarfsgerechter Beratung vor Vertragsschluss deutlich zum Ausdruck. Auch der Verhaltenskodex des Gesamtverbandes der Deutschen Versicherungswirtschaft (GDV) für den Vertrieb von Versicherungsprodukten[15] stellt unter Nr. 1 den Grundsatz auf, dass die Bedürfnisse der Kunden immer im Mittelpunkt stehen müssen.

7. Zwischenfazit

Insgesamt lässt sich festhalten, dass sich innerhalb des Versicherungssektors diejenigen Bereiche besonders für den Einsatz von KI eignen, in denen es um die Bewältigung von Routineaufgaben im sog. Massengeschäft des Versicherers mit Verbrauchern (im Gegensatz zu unternehmerisch tätigen, also gewerblichen und industriellen Versicherungsnehmern) geht. Dies beruht im Wesentlichen auf zwei Umständen, nämlich den großen verfügbaren Datenmengen und den vergleichbaren Entscheidungsgrundlagen.

III. Rechtliche Anforderungen

Im Zusammenhang mit den erwähnten Einsatzfeldern von KI im Versicherungssektor sind auch einige rechtliche Anforderungen zu beachten. Dies gilt insbesondere für die Erfordernisse des Datenschutzrechts. So gelten nach der DSGVO etwa im Bereich der Gesundheitsdaten erhöhte Anforderungen an die Verarbeitung.[16] Letztere ist gem. Art. 9 Abs. 1, Abs. 2 lit. a DSGVO grundsätzlich nur nach vorheriger ausdrücklicher Einwilligung des Betroffenen zulässig. Demgegenüber ist bei anderen Daten die Verarbeitung in vielen praxisrelevanten Fällen wie etwa dann, wenn dies zur Vertragserfüllung erforderlich ist, auch ohne eine solche Einwilligung rechtmäßig (Art. 6 Abs. 1 S. 1 lit. b–f DSGVO).

Zudem gilt es zu beachten, dass nur positive (stattgebende) Antrags- und Leistungsentscheidungen vollautomatisiert verarbeitet werden dürfen. Negative (ablehnende) Regulierungsentscheidungen scheitern an § 37 Abs. 1 Nr. 1 BDSG i. V. m. Art. 22 DSGVO, es sei denn, es handelt sich um verbindliche Entgeltregelungen in der gesetzlichen oder privaten Krankenversicherung (§ 37 Abs. 1 Nr. 2 BDSG).

Auch im Vertragsrecht sind mit dem Einsatz von KI bestimmte Anforderungen verbunden. Dabei sind einige Fragestellungen klärungsbedürftig. Dies gilt etwa für Zurechnungsfragen im Zusammenhang mit vertragserheblichen Erklärungen wie der Antragsannahme oder einer Regulierungsentscheidung. Nach der in der Diskussion vorherrschenden Ansicht sind derartige Erklärungen demjenigen zurechenbar, der den Algorithmus im Verhältnis zum Erklärungsempfänger eingesetzt

[15] https://www.gdv.de/de/themen/news/verhaltenskodex-fuer-den-vertrieb-11518.
[16] Eingehend *Waldkirch*, VersR 2020, 1141 ff.

hat.[17] Diese Einschätzung verdient Zustimmung. Dabei lässt sich das Bild des „verlängerten Arms" verwenden: Wer seinen eigenen geschäftlichen Aktionsradius erweitert, indem er – ohne selbst einzugreifen – die KI automatisierte Erklärungen abgeben lässt, muss sich diese Erklärungen zurechnen lassen. Damit ist zugleich eine Absage an Versuche verbunden, die KI als rechtsfähige, eigene Einheit oder gar als juristische Person anzuerkennen.[18] Freilich sind auch bei dem traditionellen Zurechnungsmodell einige Fragen klärungsbedürftig, etwa zur Irrtumsanfechtung bei einer Fehlfunktion des Algorithmus. Dies soll hier nicht vertieft werden.[19]

IV. Nutzen für Verbraucher

Einen Vorteil für Verbraucher kann der Einsatz von KI insbesondere im Bereich der Verwaltungskosten mit sich bringen. Dies gilt im Hinblick darauf, dass neben dem Versicherer und dem einzelnen Verbraucher als seinem Vertragspartner auch die sog. Gefahrengemeinschaft, also das Kollektiv aller Versicherungsnehmer, in den Blick zu nehmen ist.

Besonders deutlich zeigt sich dies anhand der objektiv-rechtlichen Regeln dazu, in welcher Weise die vom Versicherer erzielten Verwaltungskostenüberschüsse in der Lebensversicherung zu verteilen sind. Hierzu sieht die vom Bundesministerium der Finanzen erlassene Verordnung über die Mindestbeitragsrückerstattung in der Lebensversicherung (Mindestzuführungsverordnung) in ihrem § 8 vor, dass diese Überschüsse grundsätzlich zur Hälfte den überschussberechtigten Verträgen – und dies sind, auch wenn der Versicherer gem. § 153 Abs. 1 VVG eine solche Beteiligung abbedingen könnte, aus Wettbewerbsgründen die meisten Lebensversicherungsverträge in Deutschland – zugutekommen müssen. Darüber hinaus kommt eine Ersparnis von Verwaltungskosten auch in anderen Versicherungssparten, bei denen es an entsprechenden Vorgaben fehlt, über die Kalkulation von Prämien, Überschussbeteiligungen und Beitragsrückerstattungen regelmäßig neben dem Versicherer auch dem einzelnen Versicherungsnehmer zugute. Im Banken- und im Kapitalmarktsektor als den beiden weiteren großen Bereichen des regulierten Finanzmarkts spielt eine durch den KI-Einsatz erzielbare Kosteneinsparung im Hinblick auf die Kalkulation von Gebühren und Einlagezinsen gleichfalls eine Rolle; freilich ist der Effekt mangels eines mit Überschussbeteiligungen und Beitragsrückerstattungen vergleichbaren Systems weniger ausgeprägt als im Versicherungssektor.

---

[17] S. nur *Foerster*, ZfPW 2019, 418 (426 ff.); *Rüfner*, in: Dederer/Shin (Hrsg.), Künstliche Intelligenz und juristische Herausforderungen, 2021, 15 (22 f.).

[18] Zur Diskussion s. *Teubner*, AcP 218 (2018), 155 ff.; ablehnend etwa auch *Rüfner*, in: Dederer/Shin (Hrsg.), Künstliche Intelligenz und juristische Herausforderungen, 2021, 15 (20 f.).

[19] Näher etwa *Wendehorst/Grinzinger*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Rechtshandbuch Künstliche Intelligenz und Robotik, 2020, § 4 Rn. 71 ff.

Ein wichtiger Faktor für die Verwaltungskosten stellt der Personalaufwand dar; hinzu kommt der Raumbedarf. Insoweit kann die bereits (s. sub II 4) erwähnte KI-gestützte sog. Dunkelverarbeitung unproblematischer Informationen zu erheblichen Kosteneinsparungen führen. Außer den Kostenvorteilen vermag der KI-Einsatz hier aus Verbrauchersicht weitere Vorteile in Gestalt einer schnelleren Bearbeitung seiner Anliegen sowie – bei entsprechend qualitätvollen, auch interne Fehlerkontrollen vorsehenden Algorithmen – auch einer höheren Bearbeitungsqualität zu bieten.

Jenseits der genannten Routineaufgaben ist ein weiteres, auch aus Verbrauchersicht nutzbringendes Einsatzfeld von KI im Versicherungssektor die bereits erwähnte Betrugserkennung. Betrugsbedingte Mehrbelastungen in Gestalt ungerechtfertigter Versicherungsleistungen fließen beim Versicherer regelmäßig in die Kalkulation von Prämie, Überschussbeteiligungen und Beitragsrückerstattungen ein. Damit hat neben dem Versicherer auch die große Mehrzahl der redlichen Versicherungsnehmer die durch Versicherungsbetrug entstehenden Zusatzkosten mitzutragen. Die KI-gestützte Betrugserkennung erhöht die Richtigkeitsgewähr der Regulierungsentscheidung und hilft damit ungerechtfertigten Aufwand zu vermeiden.

Auch die oben erwähnte Kfz-Schaden-App bietet Kostenvorteile. Für den Versicherer hat diese Vorgehensweise den Vorzug, bei massenhaft auftretenden Bagatellschäden keine Personalkapazitäten für die Bearbeitung von Bagatellschäden (einschließlich der Schadensbegutachtung vor Ort) binden zu müssen, wodurch Kosten eingespart werden. Diese Ersparnis kann durch ein großzügiges, d. h. oberhalb des von dem Algorithmus errechneten voraussichtlichen tatsächlichen Schadensbehebungsaufwands liegendes Regulierungsangebot teilweise an den Versicherungsnehmer weitergegeben werden. Auf diese Weise wird es durch den Einsatz von KI ermöglicht, eine für beide Vertragspartner wirtschaftlich attraktive Abwicklung des Schadensfalls zu erzielen. Zugleich erhöht sich durch die schnelle Auszahlung einer vergleichsweise hohen Entschädigungssumme die Kundenzufriedenheit.

Im Idealfall erfolgt die Schadensbearbeitung durch KI vollumfänglich (sog. End-to-End-Bearbeitung). Bislang ist eine derartige End-to-End-Anwendung in Deutschland freilich noch nicht üblich. Vielmehr trifft die Letztentscheidung zumindest in Zweifelsfällen stets ein für die Sachbearbeitung zuständiger Mensch. Aber auch in diesem Fall lassen sich die Prozesse durch den Einsatz von KI deutlich beschleunigen.

Soweit der Nutzen des KI-Einsatzes sich in geringeren Prämien und damit für den Verbraucher in einer besseren Finanzierbarkeit des Versicherungsschutzes niederschlägt, wird damit zugleich eines der Ziele von Nachhaltigkeit verwirklicht. Begreift man diesen Begriff nämlich im weiten Sinne von Environmental Social Governance (ESG) und mithin über den Umweltaspekt hinaus, so kommt unter dem Aspekt „Social" auch der finanzierbare Versicherungsschutz zum Tragen.[20]

[20] S. dazu *Armbrüster*, ZfV 2021, 715.

V. Risiken für Verbraucher

1. Vertragsanbahnung

Bereits bei der Vertragsanbahnung kann der Einsatz von KI aus Verbrauchersicht Gefahren mit sich bringen. Als (fiktives, aber technisch durchführbares) Beispiel sei im Bereich der Antragsbearbeitung die Erkennung von Depressionserkrankungen genannt. Hier kann eine KI-gestützte Analyse anhand des Verhaltens in sozialen Netzwerken sowie der Stimmlage beim Bedienen von Sprachassistenten oder in Telefonaten zu dem Ergebnis führen, dass z. B. ein Antrag zum Abschluss einer Kranken-, Lebens- oder Berufsunfähigkeitsversicherung abgelehnt wird.

Ein derartiger Einsatz von KI erscheint – anders, als wenn etwa die Stimmanalyse zu therapeutischen Zwecken erfolgt – ethisch und rechtlich als problematisch. Was die rechtlichen Anforderungen angeht, so sind im Zusammenhang mit der Erkennung von Krankheiten datenschutzrechtliche Gebote zu beachten. Welche Regeln konkret eingreifen, richtet sich danach, ob es sich um die Verarbeitung von Gesundheitsdaten handelt. Ist dies der Fall, so darf die Verarbeitung nur aufgrund einer ausdrücklichen Einwilligung des Betroffenen erfolgen. Die Verarbeitung von Daten, die in sozialen Netzwerken kursieren, ist dann sogar gänzlich untersagt, weil sie nicht in der abschließenden Aufzählung des § 213 Abs. 1 VVG enthalten ist. Stehen hingegen nicht Gesundheitsdaten in Rede, so ist die Verarbeitung nach näherer Maßgabe der Rechtfertigungsgründe in Art. 6 DSGVO möglich. In diesem Fall bestehen deutlich niedrigere datenschutzrechtliche Hürden. So ist nach Art. 6 Abs. 1 lit. b DSGVO die Verarbeitung zulässig, wenn sie für die Erfüllung eines Vertrags, dessen Vertragspartei der Betroffene ist, oder zur Durchführung vorvertraglicher Maßnahmen erforderlich ist, und die auf Anfrage der betroffenen Person hin erfolgen.

Für die Beantwortung der Frage, ob die vom Versicherer verwendeten Daten als Gesundheitsdaten i. S. v. Art. 9 Abs. 1 DSGVO einzuordnen sind, ist die Legaldefinition in Art. 4 Nr. 15 DSGVO heranzuziehen. Demnach muss es sich um personenbezogene Daten handeln, die sich auf die körperliche oder geistige Gesundheit einer natürlichen Person, einschließlich der Erbringung von Gesundheitsdienstleistungen, beziehen und aus denen Informationen über deren Gesundheitszustand hervorgehen. Dies ist so zu verstehen, dass aus den Daten eine Information über den Gesundheitszustand der betreffenden Person hervorgehen muss. Dementsprechend handelt es sich beispielsweise nicht um ein Gesundheitsdatum, wenn der Versicherungsnehmer etwa seinen Partner oder seine Arbeitsstelle verloren hat und dieser Schicksalsschlag aus einem sozialen Netzwerk hervorgeht. Derartige Informationen lassen nämlich allenfalls mittelbar Rückschlüsse auf die Gesundheit des Versicherungsnehmers zu, und selbst dies nur in Gestalt von reinen Mutmaßungen ohne hinreichende tatsächliche Grundlage. Was Stimmdaten aus einem Telefongespräch angeht, so lassen sich zwar daraus mittlerweile tatsäch-

lich Anhaltspunkte für Depressionen und sonstige psychische Störungen ermitteln.[21] Indessen handelt es sich auch dann bei der Stimme nicht um eine direkte gesundheitsbezogene Aussage. Anders ist die Lage, wenn sich das Telefonat inhaltlich auf den Gesundheitszustand bezieht.

Betrachtet man allein den Wortlaut der zitierten Legaldefinition, so könnte man sich auf den Standpunkt stellen, dass in dem Beispielsfall mangels direkter Aussagen zum Gesundheitszustand keine Gesundheitsdaten vorliegen. Allerdings spricht Erwägungsgrund 35 der DSGVO jedenfalls tendenziell für eine weite Auslegung des Begriffs. Demnach zählen zu den personenbezogenen Gesundheitsdaten alle Daten, die sich auf den Gesundheitszustand des Betroffenen beziehen und aus denen Informationen über den früheren, gegenwärtigen und künftigen körperlichen oder geistigen Gesundheitszustand der betroffenen Person hervorgehen. Unter anderem sollen dazu auch „Informationen etwa über Krankheiten, Behinderungen, Krankheitsrisiken, Vorerkrankungen, klinische Behandlungen oder den physiologischen oder biomedizinischen Zustand der betroffenen Person unabhängig von der Herkunft der Daten […]" gehören.

Auch wenn diese Erläuterung in den Erwägungsgründen die hier zu beurteilende Fallkonstellation nicht unmittelbar betrifft, geht aus ihr doch hervor, dass der Verordnungsgeber ein weitreichendes Verständnis des Begriffs „Gesundheitsdaten" hat. Dementsprechend wird beispielsweise auch ein Portraitfoto, das eine Person mit Brille zeigt, als Gesundheitsdatum angesehen, da sich daraus eine Information über die Sehkraft des Abgebildeten entnehmen lässt.[22] Entscheidend für das Vorliegen von Gesundheitsdaten spricht der Zweck ihrer Verarbeitung durch den Versicherer. Dieser Zweck liegt darin, anhand der Daten den Gesundheitszustand der betroffenen Person zu beurteilen. Im Ergebnis ist die in dem Beispielsfall angesprochene Datenverarbeitung daher ohne ausdrückliche Einwilligung des Betroffenen unzulässig.

Einen Sonderfall stellt die Verarbeitung genetischer Daten dar. In diesem Bereich besteht aufgrund der Vorgaben des Gendiagnostikrechts für den Versicherer ein striktes Entgegennahme- und Verwertungsverbot (§ 18 Abs. 1 S. 1 Nr. 2 GenDG). Davon ausgenommen ist lediglich der Fall, dass in der Lebens-, Berufsunfähigkeits-, Erwerbsunfähigkeits- oder Pflegerentenversicherung eine Leistung von mehr als 300.000 EUR oder mehr als 30.000 EUR Jahresrente vereinbart wird. Für KI-basierte Algorithmen bedeutet dies, dass selbst dann, wenn sich aus Eingangsinformationen etwa zu Vorerkrankungen oder einer Medikamenteneinnahme Rückschlüsse auf die genetische Disposition ziehen lassen, diese Disposition als solche nicht als Kriterium für die Entscheidungsfindung herangezogen werden darf.

---

[21] https://be.invalue.de/d/publikationen/vwheute/2018/04/09/talanx-nutzt-fuer-cheffindung-software.html sowie https://www.vkb.de/content/magazin/gesundheit-medizin/e-magazin-gesundheit-aktuell/medizin/stimme/.

[22] *Ernst*, in: Paal/Pauly/Ernst, DSGVO, 3. Aufl. 2021, Art. 4 Rn. 109.

2. Schadensregulierung

Auch im Bereich der Schadensregulierung kann der Einsatz von KI für Verbraucher mit Risiken verbunden sein. Ein Beispiel bietet der Fall, dass ein Versicherer beim Einsatz der Kfz-Schaden-App die Anreizwirkung, welche die Aussicht auf eine rasche und unkomplizierte Entschädigungszahlung für viele Verbraucher bietet, dazu ausnutzt, um systematisch unter den mutmaßlichen tatsächlichen Schadensbehebungskosten liegende Beträge auszuweisen. Zwar beruht die Attraktivität der Kfz-Schaden-App nicht allein auf der raschen Schadensregulierung, sondern auch darauf, dass diese großzügig ausfällt. Daher dürfte es zu erwarten sein, dass Versicherer dieses Instrument, von dem sie auf der Kostenseite auch selbst profitieren, regelmäßig nicht zum Nachteil ihrer Kunden einsetzen werden. Allerdings wird der Verbraucher es regelmäßig ohne Hinzuziehung von Fachleuten kaum überprüfen können, ob die vom Versicherer angebotene Summe zu knapp bemessen ist. Sofern sich ein objektiv unangemessen niedriges Zahlungsangebot auf Einzelfälle beschränkt, wird man dies angesichts des auch beim Einsatz von qualitativ hochwertiger KI stets verbleibenden Restrisikos von Fehleinschätzungen und Ungenauigkeiten hinzunehmen haben. Dies gilt umso mehr, als es dem Versicherungsnehmer frei steht, das ihm durch den Versicherer unterbreitete Regulierungsangebot abzulehnen. Liegt der objektiv zu niedrigen Offerte hingegen ein systematisches Vorgehen zugrunde, so bietet dies der Aufsichtsbehörde Anlass zum Einschreiten.

Als (wiederum fiktives) weiteres Beispiel für mögliche Gefahren sei die Analyse der Klagebereitschaft von Verbrauchern angeführt. Die Frage, wie wahrscheinlich es ist, dass ein Versicherungsnehmer sich gegen eine ihm ungünstige Regulierungsentscheidung des Versicherers gerichtlich zur Wehr setzen wird, hängt in der Lebenswirklichkeit von verschiedenen Umständen ab. Zu nennen sind etwa Lebensalter und Gesundheitszustand, Bildungsstand, Beruf, geschäftliche Erfahrenheit sowie finanzielle Situation. Dementsprechend können etwa Informationen wie Gesundheitsdaten oder Kreditauskünfte gewisse Rückschlüsse erlauben. Hinzu kommen wiederum (s. bereits sub 1 zum Beispiel der Depressionserkrankung) Beiträge in sozialen Netzwerken als mögliche Informationsquelle für eine KI-gestützte Analyse der Klagebereitschaft.

Durch ein solches Vorgehen verstößt der Versicherer gegen eine im Versicherungsverhältnis als einem auf besonderem wechselseitigen Vertrauen der Vertragspartner gestützten Dauerschuldverhältnis bestehende Nebenpflicht gegenüber dem Versicherungsnehmer.[23] Hinzu kommt eine Verletzung der Datenschutzbestimmungen, da das Vorgehen nicht durch einen Rechtfertigungsgrund gedeckt ist. Auch ethisch ist das Vorgehen höchst problematisch, zumal es dem Versiche-

[23] Zum versicherungsrechtlichen Kooperationsgebot s. *Armbrüster*, Privatversicherungsrecht, 2. Aufl. 2019, Rn. 292 ff.

rungsnehmer verborgen bleibt, ob der Versicherer das Tool einsetzt, um seine Klagebereitschaft zu prüfen. All diese Einwände gelten unabhängig davon, ob die Leistungsablehnung zu Recht oder zu Unrecht erfolgt ist.

Gefahren sind nicht zuletzt auch im Zusammenhang mit der Betrugsbekämpfung zu verzeichnen. Dies gilt etwa für den Fall, dass eine Stimmanalyse vorgenommen wird, um den Wahrheitsgehalt einer telefonischen Schadensmeldung zu überprüfen. Insoweit kann sich der Versicherer nicht auf einen der in Art. 6 Abs. 1 lit. b–e DSGVO genannten Rechtfertigungsgründe stützen. Allenfalls die Wahrnehmung eigener berechtigter Interessen (Art. 6 Abs. 1 lit. f DSGVO) könnte zur Rechtmäßigkeit führen. Insoweit bedarf es freilich einer Abwägung, an die angesichts des hohen Stellenwerts, den die DSGVO dem Schutz personenbezogener Daten beimisst, strenge Anforderungen zu stellen sind.

3. Diskriminierung

Ein weiterer Themenbereich, in dem der Einsatz von KI im Versicherungssektor aus Verbrauchersicht Gefahren mit sich bringen kann, betrifft Diskriminierungen. Insbesondere bei selbstlernenden KI-Systemen droht dann, wenn in dem vom Versicherer verwendeten Algorithmus keine entsprechenden Schutzvorkehrungen getroffen werden, eine nicht gerechtfertigte und damit nach dem AGG untersagte Ungleichbehandlung. Der Abschluss und die Durchführung privater Versicherungsverträge sind im Antidiskriminierungsrecht Gegenstand eigener Regelungen in den §§ 19 Abs. 1 Nr. 2, Abs. 2, 20 Abs. 2 AGG.

Demnach gilt ein absolutes Diskriminierungsverbot (auch) im Versicherungssektor in Bezug auf das Merkmal der Rasse/ethnischen Herkunft (§ 19 Abs. 2 AGG). Zudem hat der EuGH in der Rechtssache „Test Achats“[24] entschieden, dass es hinsichtlich des Merkmals Geschlecht, auch über den (unstreitigen) Fall der Kosten für Schwangerschaft und Mutterschaft hinaus, generell keine Ungleichbehandlung geben darf. Dies gilt insbesondere auch dann, wenn Unterschiede in Prämien und Leistungen durch statistisch nachweisbare Daten begründbar sind, insbesondere hinsichtlich der höheren durchschnittlichen Lebenserwartung von Frauen. Verarbeitet nun ein Algorithmus Daten, deren Auswertung zu einer geschlechtsbezogenen Ungleichbehandlung führt, so liegt darin ein Verstoß gegen das Diskriminierungsverbot gem. §§ 19 Abs. 1 Nr. 2, 20 Abs. 2 AGG. Dafür ist eine Diskriminierungsabsicht nicht erforderlich. Die entsprechenden Daten dürfen daher nicht ungefiltert durch die KI verwertet werden.

Anders ist die Lage hinsichtlich der weiteren geschützten Merkmale Religion, Behinderung, Alter und sexuelle Identität. Insoweit lässt § 20 Abs. 2 S. 2 AGG eine Ungleichbehandlung im Versicherungssektor ausdrücklich zu, wenn sie „auf anerkannten Prinzipien risikoadäquater Kalkulation beruht, insbesondere auf einer

[24] EuGH NJW 2011, 907.

versicherungsmathematisch ermittelten Risikobewertung unter Heranziehung statistischer Erhebungen". Werden entsprechende Daten in einen Algorithmus eingespeist, so kann eine Differenzierung bei Prämien und Leistungen mithin gerechtfertigt sein.

VI. Regulierungsbedarf

Auf EU-Ebene gerät die KI-Regulierung zunehmend ins Blickfeld.[25] So hatte das Europäische Parlament zunächst in einer Entschließung Versicherungen als Systeme mit hohem Risiko einstufen wollen.[26] Dies hätte nach den Vorstellungen des Parlaments zu einer verschuldensunabhängigen Haftung für Fehlfunktionen der KI geführt.[27] Mittlerweile sieht der Entwurf der Kommission zu einer KI-Verordnung[28] nur noch eine Informationspflicht bei der Interaktion mit dem Kunden vor. Ansonsten wird für strengere Vorgaben mittlerweile ein produktbezogener Ansatz verfolgt, der etwa die Bereiche Spielzeug, Medizinprodukte und Kfz in den Blick nimmt.

Ein Bedarf an Regulierung des Einsatzes von KI im Versicherungssektor lässt sich im Wesentlichen in drei Bereichen feststellen, nämlich Transparenz, Qualitätssicherung und Kontrolle. Was die *Transparenz* angeht, so kommt Informationspflichten des Versicherers gegenüber dem Versicherungsnehmer eine zentrale Rolle zu.[29] Das Gebot der Nachvollziehbarkeit von KI-Algorithmen[30] ist insbesondere im Hinblick auf die Kontrolle durch die Aufsichtsbehörden bedeutsam. Unter dem Schlagwort „Explainable AI" finden gegenwärtig verstärkt Anstrengungen statt, die Nachvollziehbarkeit zu verbessern und damit zugleich die Akzeptanz des KI-Einsatzes durch Verbraucher zu erhöhen.[31]

---

[25] Überblick bei *Hacker*, NJW 2020, 2142 ff.

[26] Entschließung des EU-Parlaments v. 20.10.2020, P9_TA(2020)0276; https://versicherungswirtschaft-heute.de/maerkte-und-vertrieb/2020-11-23/eu-parlament-stuft-einsatz-von-ki-bei-versicherern-als-hochriskant-sein/ – S. dazu *Etzkorn*, CR 2020, 764 ff.

[27] Vgl. dazu das 10-Punkte-Papier des GDV; https://www.gdv.de/de/themen/politische-positionen/stellungnahmen/ergaenzende-ki-regulierung-nur-fuer-hochriskante-anwendungen-64528.

[28] KOM(2021) 206 final, https://eur-lex.europa.eu/legal-context/DE/ALL/?uri=CELEX:52021PC0206.

[29] S. dazu allg. den informationsbezogenen Ansatz der EU-Kommission in ihrem Entwurf zu einer KI-Verordnung, KOM(2021) 206 final, https://eur-lex.europa.eu/legal-context/DE/ALL/?uri=CELEX:52021PC0206.

[30] https://www.bitkom.org/Bitkom/Publikationen/Blick-in-die-Blackbox-Nachvollziehbarkeit-von-KI-Algorithmen-in-der-Praxis.

[31] S. etwa https://targens.de/news/explainable-ai-wie-ki-finanzinstitute-im-compliance-bereich-unterstuetzen-kann/ – Zum derzeit noch geringen Vertrauen in KI-gestützte Anwendungen s. https://home.kpmg/de/de/home/themen/2021/05/studie-buergerinnen-und-buerger-haben-wenig-vertrauen-in-ki.html.

Die *Qualität* KI-gestützter Entscheidungen lässt sich durch eine ausreichende Trainingsphase verbessern. Hinzu kommt die Einhaltung von Anforderungen des Datenschutzes und der IT-Sicherheit. Zu erwägen ist zudem eine bundes- oder europaweite Datenplattform[32], um den Algorithmen eine möglichst breite Datenbasis verfügbar zu machen.

Was schließlich die *Kontrolle* angeht, so führen die oben genannten Beispielsfälle deutlich vor Augen, welche gravierenden Gefahren aus Verbrauchersicht mit dem Einsatz von KI im Versicherungssektor verbunden sein können. Hier bietet eine Aufsicht der BaFin über die von Versicherern eingesetzten Algorithmen die Möglichkeit Fehlentwicklungen präventiv gegenzusteuern. Die BaFin hat bereits im Jahr 2018 hierzu eine Studie mit dem Titel „Big Data trifft auf künstliche Intelligenz“[33] veröffentlicht, in der zentrale Themen angesprochen werden. Insoweit kommt auch eine Anmeldeverpflichtung der Versicherer für KI-Anwendungen, deren Einsatz sie beabsichtigen, in Betracht. Ein weiteres Instrument ist ein noch zu schaffendes Gütesiegel in Gestalt einer KI-Zertifizierung, bei deren Verleihung neben rechtlichen auch ethische Aspekte berücksichtigt werden sollten.[34]

VII. Fazit und Ausblick

1. Chancen und Risiken

Der Einsatz von KI im Versicherungssektor bietet aus der Verbraucherperspektive manche Chancen. Insbesondere lässt sich der Zugang zum Versicherungsschutz erleichtern, indem Kostenersparnisse erzielt werden, welche auch dem Kollektiv und damit letztlich jedem einzelnen Versicherungsnehmer zugutekommen. Auch können zuverlässige und zügige Regulierungsentscheidungen zu einer erhöhten Kundenzufriedenheit führen sowie Rechtsstreitigkeiten vermeiden. Zugleich birgt der KI-Einsatz aber auch in verschiedener Hinsicht Risiken. Ihnen gilt es mit den Instrumenten des Datenschutzrechts und der Legalitätskontrolle durch die Versicherungsaufsicht entgegenzutreten.

Von zentraler Bedeutung dafür, dass Versicherungsnehmer gegenüber dem Einsatz von KI aufgeschlossen sind und angesichts der damit verbundenen Vorzüge auch die erforderlichen datenschutzrechtlichen Einwilligungen erteilen, ist Vertrauen. Dieses lässt sich durch eine transparente Handhabung des Instruments gewinnen. Zudem sollte der Einsatz von KI nicht allein den Interessen und Zielen des Versicherers dienen, sondern von vornherein auch auf einen Mehrwert für

---

[32] Vgl. zum Projekt GAIA-X https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html.

[33] https://www.bafin.de/SharedDocs/Downloads/DE/dl_bdai_studie.html.

[34] Vgl. etwa auch die IT-bezogenen Angebote von DEKRA (https://www.dekra.de/de/cybersecurity/) oder TÜV (https://www.tuvit.de/de/startseite/).

die Versicherungsnehmer ausgerichtet sein. Damit können Win-Win-Situationen geschaffen werden. Das obige Beispiel der Schadensregulierung über eine Kfz-Schaden-App belegt dies eindrucksvoll.

2. Ethische Anforderungen

Nicht zuletzt sollte der Einsatz von KI auch ethischen Grundanforderungen gerecht werden.[35] Das soeben im Zusammenhang mit der Vertrauensbildung aufgestellte Postulat, dass Versicherer KI nur mit Mehrwert für die Kundschaft einsetzen sollten, lässt sich auch als ethische Forderung begreifen. Der in dem oben (sub II 6) erwähnten Verhaltenskodex des GDV für den Vertrieb von Versicherungsprodukten aufgestellte Grundsatz, dass die Bedürfnisse der Kunden immer im Mittelpunkt zu stehen haben, ist insoweit sinngemäß auch für den Einsatz von KI im Versicherungssektor übertragbar.

Hinzu kommt als weiteres ethisches Gebot, dass selbst bei einem weitreichenden KI-Einsatz stets auch ein menschlicher Faktor erhalten bleiben sollte. Der europäische Gesetzgeber hat angeordnet, dass eine vollautomatische Entscheidung im Einzelfall unter bestimmten Voraussetzungen zulässig sein kann (Art. 22 Abs. 2 DSGVO). Freilich sollte darüber aus ethischer Sicht hinausgegangen und – jedenfalls jenseits von Bagatellfällen – die regelhafte Einschaltung eines Menschen beibehalten werden.

Dies gilt beispielsweise für den Umgang des Versicherers mit Kulanz. Zwar lassen sich viele der objektiven Faktoren, die für eine Kulanzentscheidung maßgeblich sein können, wie etwa Dauer und Umfang der Vertragsbeziehung, bisheriger Schadensverlauf und wirtschaftliche Verhältnisse des Versicherungsnehmers, auch durch Algorithmen abbilden. Gleichwohl bleibt bei der Schadensregulierung meist ein gewisser Ermessensspielraum in Bereichen, die sich einer rein rationalen Erfassung entziehen. Dies gilt etwa dann, wenn der tatsächliche Geschehensablauf nicht vollständig aufgeklärt werden kann und die Glaubhaftigkeit von Angaben des Versicherungsnehmers in Rede steht. Auch weitere Faktoren wie die sonstigen Auswirkungen eines als Versicherungsfall gemeldeten Schadensfalls auf den Versicherungsnehmer können insoweit eine Rolle spielen. Man könnte sich auf den Standpunkt stellen, dass der Einsatz von KI hier zu einer sachgerechten Rationalisierung der Regulierungsentscheidung führt. Indessen kann gerade in solchen Situationen dann, wenn ein Mensch die Lage beurteilt und zugunsten des Versicherungsnehmers entscheidet, die Kundenbindung verstärkt werden. Zudem kann von diesem menschlichen Faktor nicht nur der Ruf des Versicherers profitieren, sondern auch derjenige der Versicherungswirtschaft insgesamt mit ihrer gesellschaftlichen Verantwortung.

---

[35] S. dazu allg. im Kontext von KI-Anwendungen *Möslein*, RD 2020, 34 ff.; *Schliesky*, NJW 2019, 3692 ff.

The AI Act Proposal: Towards the next transparency fallacy?

Why AI regulation should be based on principles based on how algorithmic discrimination works

Bettina Berendt[1]

I. Introduction

Artificial Intelligence (AI) can entail large benefits as well as risks. The goals of protecting individuals and society and establishing conditions under which citizens find AI "trustworthy" and developers and vendors can produce and sell AI, the ways in which AI works have to be understood better and rules have to be established and enforced to mitigate the risks. This task can only be undertaken in collaboration. Computer scientists are called upon to align data, algorithms, procedures and larger designs with values, 'ethics' and laws. Social scientists are called upon to describe and analyse the plethora of interdependent effects and causes in socio-technical systems involving AI. Philosophers are expected to explain values and ethics. And legal experts and scholars as well as politicians are expected to create the social rules and institutions that support beneficial uses of AI and avoid harmful ones.

This article starts from a computers-and-society perspective and focuses on the action space of lawmaking. It suggests an approach to AI regulation that starts from a critique of the European Union's (EU) proposal for a Regulation commonly known as the AI Act Proposal, published by the EU Commission on 21 April 2021.[2]

---

[1] I thank Laurens Naudts, Geoffrey Rockwell, Pieter Delobelle, Rainer Rehak, Kristen Scott and Koen Vraenckaert for their helpful comments on an earlier version of this work and the participants of the conferences "Rechtliche Rahmenbedingungen für KI in der Schweiz" (Zürich/online, November 2021) as well as "Verbraucherrechtstage 2021" (online, July 2021) for important discussions. I received funding from the German Federal Ministry of Education and Research (BMBF) – Nr. 16DII113 f. I am also indebted to the inputs from the projects VeriLearn (Research Foundation – Flanders (FWO), EOS No. 30992574) and NoBIAS (NoBIAS – H2020-MSCA-ITN-2019 project GA No. 860630).

[2] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM(2021) 206 final 2021/0106 (COD), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

The AI Act Proposal deals with a wide range of phenomena that have been and are being discussed in large bodies of literature, including risks deemed unacceptable, manipulation, and health and safety risks. This paper will concentrate on phenomena related to AI, algorithmic systems and Big Data, discussed under terms such as "discrimination", "fairness" or "bias". I argue that regulation should be based on science and that, viewed from this angle, the AI Act Proposal falls short. In particular, its strong focus on transparency as a measure against bias and discrimination ignores relevant research on how algorithmic discrimination 'works' and how it can be countered.[3]

The AI Act Proposal is being discussed from many angles, in academic, political, and other fora. The present paper was motivated, in particular, by the analysis of *Veale* and *Zuiderveen Borgesius*,[4] but differs from it in its thematic focus (bias and discrimination) and in the approach it takes (the search for principles). Like EDRi et al.,[5] I consider it key to ground AI regulation in the protection of fundamental and human rights.

The article is organized as follows. In Section II, key terms are defined. Section III highlights the role of algorithmic discrimination in the AI Act Proposal and in the General Data Protection Regulation (GDPR) and outlines the relationship between discriminatory effects via the processing of personal data and via AI-based processing. Section IV takes a closer look at how algorithmic discrimination arises; in particular, it emphasizes the cumulative effects of human and machine biases/discrimination in real-life chains of processes, using labour as an example domain. Section V argues that the AI Act Proposal claims that the measures it requires are suitable and that these measures rely centrally on transparency. This assumption is challenged in Section VI, which shows that transparency, even if coupled with data quality, security, human oversight and documentation, is not sufficient for the goal of preventing discrimination (and in this sense, if relied upon as the core requirement, not suitable). Section VII proposes to build on the construction of data protection on a web of principles (of which transparency

---

[3] The Explanatory Memorandum points out that the proposal "is the result of extensive consultation with all major stakeholders" (p. 7 of COM[2021] 206 final 2021/0106 [COD]), "state-of-the-art for many diligent operators", "derived from the Ethics Guidelines of the HLEG, piloted by more than 350 organisations", and "largely consistent with other international recommendations and principles" (p. 13). It also notes that "[t]he precise technical solutions to achieve compliance with those requirements may be provided by standards or by other technical specifications or otherwise be developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the AI system" (p. 13). While this refers to the science and may be indirectly influenced by the science, it stops short of constituting a scientifically based set of norms and rules.

[4] *Veale/Zuiderveen Borgesius*, Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, 2021, 22 (4), 97–112.

[5] EDRi (European Digital Rights) et al., An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement, https://epicenter.works/sites/default/files/political-statement-on-ai-act.pdf, 2021.

is one, but not the only one) to shape AI regulation. Section VIII concludes by formulating legal and computer-science goals for better AI regulation: legal principles and software engineering recommendations that map principles to design strategies and these to design patterns and technologies.

II. Terminology: (Algorithmic) bias and (algorithmic) discrimination

(Unlawful) discrimination consists in making differentiations on the basis of objectionable or illegal grounds, for example on the basis of gender, sexual preference, or ethnic origin.[6] *Algorithmic discrimination (AD)* can be defined as discrimination in contexts that involve (usually digital) computers. *Friedman* and *Nissenbaum* argued, against a then frequent conception of computers as 'more objective' than humans, that computer systems can in fact be biased and can lead to discrimination. They "use the term *bias* to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others".[7]

On the one hand, their definition overlaps with the legal notion: "A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate".[8] On the other hand, this definition allows for any type of grounds of the differential treatment, including for example "effecting a long/resource-intensive computing job in a multi-user computer system". Thus, *Friedman* and *Nissenbaum* make no a priori commitment regarding whether the differentiation is unfair (in a moral sense) and whether it constitutes discrimination (in a legal sense).

AD can amount to direct discrimination, but also and more typically to indirect discrimination.[9]

Discrimination and AD are linked to complex questions of justice, equality, and fairness. In line with the dominant terminology in the current machine-learning literature and to simplify, out of these three only "fairness" will be used here.

Berendt gives an introduction to the computational-legal discussion of these questions.[10] The examples of AD in Section IV concentrate on ethnicity and gen-

---

[6] *Zuiderveen Borgesius*, Strengthening legal protection against discrimination by algorithms and artificial intelligence, The International Journal of Human Rights, 2020, 24 (10), 1572–1593.

[7] *Friedman/Nissenbaum*, Bias in computer systems, ACM Transactions on Information Systems, 14(3), 1996, 330–347, 332.

[8] *Friedman/Nissenbaum* (Fn. 7), 332.

[9] E. g. *Pedreschi/Ruggieri/Turini*, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York City (NY), 2008, 560–568; *Barocas/Selbst*, Big data's disparate impact, California Law Review, 104(3), 2016, 671–732.

[10] *Berendt*, Algorithmic discrimination, in: Comandé (Ed.), Elgar Encyclopedia of Law and Data Science, Cheltenham (UK), 2022.

der that are legally protected attributes in many jurisdictions. The attribute (body) weight will be used in one example to demonstrate how the openness of the AD definition above allows us to also consider a wider range of grounds.[11] All of these attributes are typically personal data, so the question arises whether data protection law might already be sufficient to counter AD.

III. Algorithmic systems and discriminatory effects: GDPR and AI Act Proposal

The AI Act is commonly regarded as being part of a 'family' of recent EU legislation, some of which is already in effect (GDPR) and some of which is currently in various stages of deliberation (Data Governance Act, Digital Services Act, Digital Markets Act, the updated General Product Safety Regulation, and the Product Liability Directive). The GDPR is particularly relevant, since discriminatory effects are often linked to the processing of personal data. Data protection and anti-discrimination are linked and convergent also in further legal[12] and computational[13] ways. In this section, I will survey the ways in which the GDPR and the AI Act Proposal address AD, explain why many but not all discriminatory effects are covered by the GDPR, and highlight core potentials and limitations of the two laws.

The GDPR requires personal data to be processed *in such a way as to* not produce discriminatory effects (Recitals 71, 75 and 85). The AI Act Proposal motivates its regulatory approach, inter alia, with reference to discrimination: *because* certain AI systems can have discriminatory effects, they should be operated only under constraints or be forbidden altogether (Recitals 15, 17, 28, 33, 35–39, 44 and 47). Thus, both laws recognize that AD is a specific and relevant risk of algorithmic systems (usually AI systems working on big personal data and similar software-based systems[14]) and that this risk often arises from the processing of per-

---

[11] Weight is an attribute that some argue should become a legally protected attribute, cf. *Schallenkamp/DeBeaumont/Houy*, Weight-Based Discrimination in the Workplace: Is Legal Protection Necessary?, Employee Responsibilities and Rights Journal, 2012, 24 (4), 251–259; *McCall/Bever*, Current trends in combating weight discrimination in the workplace, https://www.fisherphillips.com/news-insights/current-trends-in-combating-weight-discrimination-in-the-workplace.html, 2020.

[12] *Gellert/de Vries/de Hert/Gutwirth*, A comparative analysis of anti-discrimination and data protection legislations, in: Custers/Calders/Schermer/Zarsky (Eds.), Discrimination and Privacy in the Information Society. Data mining and Profiling in Large Databases, Berlin etc. 2013, 61–89; *Naudts*, How machine learning generates unfair inequalities and how data protection instruments may help in mitigating them, in: Leenes/van Brakel/Gutwirth/de Hert (Eds.), Data Protection and Privacy: The Internet of Bodies, Oxford 2019, 71–92.

[13] *Berendt* (Fn. 10).

[14] This argument follows the German Data Ethics Commission's terminological and semantic proposal to not limit attention to "AI" in any specific technical sense, Datenethikkommission, Opinion of the Data Ethics Commission, https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3, 2019.

sonal data. These claims are in line with a body of literature that has been growing strongly since the mid-1990s.[15]

However, the question arises whether this is enough.

To the extent that the laws talk about discrimination, they implicitly refer to the applicable notion(s) of discrimination in anti-discrimination law. However, neither of the two laws defines discrimination or bias in light of the specific particularities and challenges of AD, and neither stipulates specific principles or measures against them. This, I argue, does not take sufficient advantage of the available science. In the remainder of this article, I will survey pertinent findings and from these derive recommendations for regulation.

At this point, it is important to consider possible reasons for these gaps, for the differences between the gap in the AI Act Proposal from that in the GDPR, and to ask what we could learn from the GDPR for the AI Act Proposal.

Algorithmic systems can (a) involve (the processing of) personal data, (b) violate data protection law, and/or (c) have discriminatory effects. All combinations are possible (except combinations with (b) + not (a), since (b) requires the involvement of personal data), as the following examples show.

i. Discriminatory effects + violating data protection law (or at least data protection principles): This constellation is in the focus of many current debates around the business models of large online platforms. Via the large-scale collection of personal data on their users (what people posted, clicked on, interacted with, what friends they have and how these behave, …), advertising can become highly personalized. Marketers who buy advertising space (and the automated, real-time online bidding algorithms that control the serving of ads) often do so based on the demographics of their target groups – and platforms have such data about each individual user. The data may be true or fictitious (an individual may have an online identity with, say, a self-specified gender that deviates from the one they have or use otherwise); the data may be given explicitly or inferred by a software; and the matching may be done inside and by the platform (probably legally) or the data may be sold to the advertising partners (probably violating GDPR provisions – see the 2022 ruling of the Belgian Data Protection Authority against IAB's Transparency and Consent framework[16]). This may lead, for example, to women systematically being shown ads for lower-paying jobs than men (see Section IV).

---

[15] See *Berendt* (Fn. 10) for a survey and discussion. Seminal articles include *Friedman/Nissenbaum* (Fn. 7), *Pedreschi/Ruggieri/Turini* (Fn. 9) and *Barocas/Selbst* (Fn. 9). Ongoing conference and workshop series contribute to the further development of the field, e. g. https://www.facctconference.org/network.

[16] Decision on the merits 21/2022 of 2 February 2022, Case number: DOS-2019-01377, English translation available at https://www.gegevensbeschermingsautoriteit.be/publications/beslissing-ten-gronde-nr.-21-2022-english.pdf.

Such personalization ("targeting") is increasingly being viewed with suspicion, and large platforms are promising to abandon certain forms of it.[17]

ii. Discriminatory effects + involving personal data but compliant with data protection law and principles: In *Heinz Huber v Germany*[18], the European Court of Justice found that while a registry of personal data kept by German authorities complied with data protection law under specified restrictions, it was discriminatory because it treated non-German EU nationals different from German citizens. Examples more closely linked to AI can be found in Section IV.

iii. Discriminatory effects without involvement of personal data: A machine-learning algorithm that learns a classifier from data that reflect past discriminatory patterns, can replicate these patterns. The learning may happen on personal data (such as microdata) or on anonymous data or even on synthetic data (generated on the basis of statistical distributions on real-life data). The last two types of training data are not considered personal data. The application of the classifier to new individuals, for example in order to decide on allocations of goods and services (such as job ads) or decisions with significant consequences such as recruitment decisions involves personal data (namely those of the new individual).

However, harms may also arise without any involvement of such data in algorithmic processing. Harms can arise through biases, stereotypes and similar that are implicit in text corpora and that can have subtle, unexpected, and still pervasive effects in subsequent data processing steps, as described in more detail in Section IV below.

Both laws are motivated by the goal of avoiding risks, but their starting points differ.

Data protection law is, to a large extent, driven by the insight that structural asymmetries, often power asymmetries, pose risks to fundamental rights. The GDPR emphasises, in various places, that it considers "risks to the fundamental rights and freedoms" – the rights to data protection and privacy in particular, but also all others. This implies that the GDPR is also meant to, and designed to, protect against risks of individual discrimination. The AI Act Proposal, on the other hand, is strongly patterned on product safety laws.[19] Thus, products (now extended to also include AI systems) are conceived of as sources of risks and harms, and a process resulting in a CE mark for an AI system[20] is proposed as a remedy.

---

[17] E. g. The Guardian, Facebook bans ads targeting race, sexual orientation and religion, https://www.theguardian.com/technology/2021/nov/10/facebook-bans-ads-targeting-race-sexual-orientation-and-religion, 2021.

[18] Case C-524/06 Heinz Huber v Bundesrepublik Deutschland [2008] ECR 2008 I-09705.

[19] *Veale/Zuiderveen Borgesius* (Fn. 4).

[20] See the definition in Article 3 (24) AI Act Proposal. A commercial product with a CE mark ("conformité européenne") indicates that the manufacturer or importer affirms the good's conformity with European health, safety, and environmental protection standards. The CE marking is required for goods sold in the European Economic Area (EEA), but is also found on products sold elsewhere that have been manufactured to EEA standards.

Both laws also are aware of societal risks, but both 'lineages' can create blind spots and challenges for protecting against social risks.

Data processing affects not only the individual, but also collectives of various kinds, up until and including the fabric of democracy itself.[21] Thus, increasingly, data protection law has to be interpreted also with a view to protecting against societal harms. However, the GDPR's focus on (individual) fundamental rights presents challenges; for example, the effects of profiling on social groups are difficult to subsume.[22]

The AI Act Proposal explicitly regulates products (AI systems) that lead to "detrimental or unfavorable treatment of certain persons or whole groups thereof" (Article 5 (1) (c)), but it does so in the relatively narrow context of the prohibited AI systems regulated in Article 5. Challenges for dealing adequately with these risks derive from the AI Act Proposal's heritage from product safety, which is typically the protection of an (individual) 'user' of a product against risks that arise from the interaction between said product and said individual, risks that in traditional products are handled by requiring certain modifications to those products. Therefore, the AI Act Proposal requires modifications to these AI systems rather than taking a wider view of the algorithmic systems / information systems / socio-technical systems that the AI is part of and that are often the underlying origins of risks and harms, such that AI primarily 'automates [existing] inequality' (Eubanks, 2018).

Both laws operate in a socio-legal context in which discrimination itself is predominantly viewed as deriving from a locatable decision. The traditional focus on individual human decision-makers who discriminate has led to a corresponding focus on individual machine decision-makers that discriminate[23]. As a result, the law and its enforcement face difficulties in the attempt to protect against structural discrimination. This problem has been recognised with regard to discrimination in the workplace under changing managerial policies, even before the advent of AD.[24]

---

[21] See the German Constitutional Court's Census Judgment of 1983: "Persons who assume, for example, that attendance of an assembly or participation in a citizens' interest group will be officially recorded and that this could expose them to risks will possibly waive exercise of their corresponding fundamental rights (Articles 8 and 9 of the Basic Law). This would not only restrict the possibilities for personal development of those individuals but also be detrimental to the public good since self-determination is an elementary prerequisite for the functioning of a free democratic society predicated on the freedom of action and participation of its members." (Bundesverfassungsgericht [BVerfG], Urteil vom 15.12.1983 – 1 BvR 209/83, 1 BvR 269/83, 1 BvR 362/83, 1 BvR 420/83, 1 BvR 440/83, 1 BvR 484/83, translation at https://freiheitsfoo.de/census-act/, retrieved 2021-12-07).

[22] Taylor/Floridi/van der Sloot (Eds.), Group Privacy: new challenges of data technologies, Dordrecht 2017.

[23] As an example, consider the wording of Article 22 GDPR.

[24] *Bagenstos*, The structural turn and the limits of antidiscrimination law, California Law Review, 94(1), 2006, 1–47.

The present paper starts from a focus on risks for fundamental rights and for society, and from the assumption that data protection's roots in tracing problems back to structural (power) asymmetries can help us re-centre structural asymmetries and structural discrimination when regulating AI. As the discussion in the present section has shown, the GDPR alone is not sufficient to address these issues.

In the next section, we will take a closer look at examples of discrimination that for the most part are not directly linked or linkable to a specific use of an individual's personal data or to a specific decision that can be called discriminatory.

IV. How and why do algorithms discriminate?

AD arises from interactions between data and algorithms and the larger socio-technical systems they are used in. To illustrate these interactions, I will sketch computational effects with the help of examples from the scientific literature and popular media.

First, we will consider how AD can result from key elements of machine learning – the training data, the algorithms, and category labels – in an idealized linear pipeline of functioning.

When data are biased, so will system representations learned from these data (without corrections). The Twitter bot Tay, designed to learn to tell jokes from interacting with human Twitter users, learned to spew out racial slurs because it was fed with racist input. This can happen very fast – Tay bot was taken offline within a day.[25] In predictive policing, using data on drug-related arrests for training has led to system recommendations to patrol the area in which most arrests were made, leading to more arrests in this area. When these areas are predominantly 'black' neighbourhoods, drug use in 'white' areas becomes under-patrolled and under-reported.[26] In addition, the statistical under-representation of some demographics in training data can lead to more prediction errors. This has been observed for facial recognition algorithms that have led to several innocent black men being arrested.[27]

Machine-learning algorithms need to have some strategy for generalising beyond the data they have seen, i. e. in order to be able to function at all. This can lead to socially biased algorithmic output. For example, recommender algorithms that are based on popularity can lead to people receiving sexist, racist, etc. claims as

---

[25] *Wakefield*, Microsoft chatbot is taught to swear on Twitter, BBC News, 2016, https://www.bbc.com/news/technology-35890188.

[26] *Lum/Isaac*, To predict and serve? Significance, 2016, 13(5), 14–19.

[27] *Hill*, Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. The New York Times, https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html, 2020.

query-completion suggestions of what they may be looking for[28] or the recommendation to consult somebody's criminal record when they search for a black-sounding name more often than when searching for a white-sounding name[29]. Algorithms designed to detect and filter out pornographic images have been alleged to compare the percentage of nude-coloured pixels with a threshold, which may lead to disproportionately high (mis)classifications for images of overweight people, which in turn can lead to disproportionate blocking of content and accounts.[30]

The categories to which a predictor maps can be a source of further problems. For example, body scanners at airports are designed to detect 'unexpected' shapes and objects on passengers – which means they have to 'expect' for example bra underwiring on female passengers, or genitals on men, to avoid causing an alarm for every single person. Airport staff therefore inform the body scanner, usually by the press of a button, of the passenger's gender before scanning. This applied binary gender schema can lead to patterns of false alarms and disproportionately burden non-binary people with pat-downs and further questions.[31]

Second, we will consider the combined effects of several steps of machine and human learning. Such combinations can produce strong AD effects. The cumulation can be understood using the notions of allocative and representational harm.

AD can consist in withholding opportunities or resources (*allocative harm)*,[32] or in perpetuating stereotypes and cultural denigration (*representational harm*). In the examples above, many harms are allocative (arrests, account blocking, airport security treatment), while some seem primarily representative (jokes, recommendations). Discrimination often works through chains of such harms, an effect that has been termed "pernicious" or "runaway" feedback loops.[33] Such loops have been described as occurring in environments dominated by one learning algorithm (ibid.) or in multi-algorithmic environments such as the Web.[34] They can

---

[28] UN Women, UN Women ad series reveals widespread sexism, http://www.unwomen.org/en/news/stories/2013/10/women-should-ads, 2013.

[29] *Sweeney*, Discrimination in Online Ad Delivery. Communications of the ACM, 2013, 56(5), 44–54.

[30] *Richman*, This is the impact of Instagram's accidental fat-phobic algorithm. https://www.fastcompany.com/90415917/this-is-the-impact-of-instagrams-accidental-fat-phobic-algorithm, 2019.

[31] *Waldron*/*Medina*, When transgender travelers walk into scanners, invasive searches sometimes wait on the other side, ProPublica, https://www.propublica.org/article/tsa-transgender-travelers-scanners-invasive-searches-often-wait-on-the-other-side, 2019.

[32] *Blodgett*/*Barocas*/*Daumé III*/*Wallach*, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg / PA 2020, 5454–5476; also referred to as *distributive harm*, *Binns*, Fairness in Machine Learning: Lessons from Political Philosophy, in: Proceedings of Machine Learning Research 81, 2018, 149–159.

[33] *O'Neil*, Weapons of Math Destruction, New York 2016; *Ensign*/*Friedler*/*Neville*/*Scheidegger*/*Venkatasubramanian*, Runaway Feedback Loops in Predictive Policing, Proceedings of Machine Learning Research 81, 2018, 160–171.

[34] *Baeza-Yates*, Bias on the web, Communications of the ACM 61(6), 2018, 54–61.

become even more pernicious when many different, independent actors and steps, including non-algorithmic ones, are involved. I will illustrate this with a cumulation scenario of phenomena observed around algorithms involved in people's paths into jobs. Each step is annotated with its predominant type of discrimination, type of harms, and involvement of personal data / data protection violations.

1. People (prospective candidates as well as prospective employers) perceive a world in which high-paying, prestigious jobs tend to be held by men (and less prestigious jobs by women).

This perception may derive from actual observation, but more frequently it is mediated through the texts one reads – but algorithms can mirror and exacerbate imbalances. This is easy to see in machine translation. For example, the German sentence "Sie ist eine gute Ärztin" (she is a good doctor) was translated, by Google Translate[35], to Turkish as "O iyi bir doktor", but then back to German as "Er ist ein guter Arzt" (he is a good doctor). This is correct in the sense that Turkish has no gender pronouns like German or English, but biased in the sense that the machine translation chooses the male interpretation.

This bias is probably due to the much higher frequency of example sentences found in parallel corpora online (= the training data of machine translation algorithms) in which the doctors described were actually male.[36] Google has been called out on such biases and addressed them in its translations to and from English: "O iyi bir doctor" is translated as "he is a good doctor *or* she is a good doctor". However, this improvement is an exception local to English, and it is brittle. For example, "Murat is her son" gets translated as "Murat onun oğlu" and back as "Murat is his son".

[Direct discrimination; representational harm; personal data generally not involved]

2. Women are served ads for different jobs than men. *Datta* et al. found that high-paying jobs tended to be served to men.[37] *Kayser-Bril* showed, in an experimental study, that Facebook served specific job ads mostly to men or mostly to women (e. g. machine learning developer vs. nurse), even if the ads are phrased in gender-neutral ways.[38] *Imana*, *Korolova*, and *Heidemann* found similar bias for *one* job type (delivery driver) mirroring the gender ratios in the respective company that placed the ad.[39]

---

[35] All translations were generated in July 2021.

[36] This interpretation can be substantiated by using a translation engine that shows context sentences, such as context.reverso.net.

[37] *Datta/Tschantz/Datta*, Automated experiments on ad privacy settings, Proceedings on Privacy Enhancing Technologies 2015 (1), 92–112.

[38] *Kayser-Bril*, Automatisierte Diskriminierung: Facebook verwendet grobe Stereotypen, um die Anzeigenschaltung zu optimieren, https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren/, 2021.

[39] *Imana/Korolova/Heidemann*, Auditing for discrimination in algorithms delivering job ads, in: Proceedings of The Web Conference 2021 (WWW '21), New York City (NY) 2021, 3767–3778.

[Direct discrimination; allocative harm; personal data are involved, currently not considered a data protection law violation[40]]

3. Ads are phrased in ways that women tend not to apply for a position because "they do not find themselves in it" or because of lower self-esteem.[41]

[Direct discrimination operating through psychological effects; can lead to allocative harm; no personal data involved]

4. Descriptions of jobs and descriptions of people holding them are machine-learned as describing good matches of CVs to jobs. Through the historical over-representation of men in these jobs, the machine may learn to map attributes that are typically found in job applications written by women to the class of not suitable applicants. The algorithm may pick up the explicit gender attribute as a predictor; more frequently though, it will choose 'female wording' (see previous item) or even all-women educational institutions as predictors.[42] Business goals affect what prediction errors the algorithm is designed to avoid, which can lead to further bias.[43] As a result, fewer women will be invited to job interviews, and fewer women will be recruited.

[Indirect discrimination; allocative harm; applicant's personal data involved]

5. The effects in 4. may be exacerbated when modern pre-processing methods for language understanding are used that rely on word embeddings or pre-trained language models. Through the biases embedded in the large corpora on which these intermediate models are learned,[44] associations with skills, characteristics, and features that may influence the subsequent classification in subtle ways, can be learned and may lead to biased outcomes.[45]

[Indirect or direct discrimination; representational harm; personal data may or may not[46] be involved.]

---

[40] Facebook describes targeting by gender as a technically possible and legitimate choice: https://www.facebook.com/business/help/151999381652364 (retrieved 21 Dec. 2021).

[41] *Burell*, Die Diskriminierung steckt oft im Detail, Der Spiegel, 15 June 2021, https://www.spiegel.de/start/stellenanzeigen-werden-oft-fuer-maenner-formuliert-wie-frauen-trotzdem-den-job-bekommen-a-2ff0215c-009c-48b1-b045-a462ca808cd7.

[42] *Dastin*, Amazon scraps secret AI recruiting tool that showed bias against women, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G 2018; *Lauret*, Amazon's sexist AI recruiting tool: how did it go so wrong?, https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e, 2019.

[43] *Lauret* (Fn. 42).

[44] Of the type 'doctors are men, nurses are women', *Bolukbasi*/*Chang*/*Zou*/*Saligrama*/*Kalai*, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, 4349–4357.

[45] Specific complications arise in languages that are more strongly gendered grammatically than English, for example Dutch (*Delobelle*/*Winters*/*Berendt*, RobBERT: a Dutch RoBERTa-based Language Model, in: EMNLP [Findings] 2020: 3255–3265).

[46] Aggregate statements ("women are …" etc.) usually do not involve personal data.

6. Public employment services, a different and independent actor than those in 1.–5., also rely on data analysis. Due to historically grown imbalances on the labour market, machines learn that women are less likely to find (re-)employment fast and therefore assign them a worse risk score/class in a statistically based risk-assessment system. If this classification leads to the job-seeker not receiving funding for (re-)training courses, they may find it harder still to find a new position, and their self-esteem will suffer. Due to the results of classification, this will affect women more often.

Additional effects may result from variable and user interface design. In the system analysed by *Allhutter* et al., only female job-seekers are asked whether they have "care obligations" and whether these are "being taken care of" such that they can be available fully to the labour market.[47] Since this variable is lacking for men, any predictive power of this attribute will likely discriminate against some women.

[Indirect discrimination, regression-based models may also lead to direct discrimination; allocative and representational harm; personal data of the new job seeker involved, whether data protection rights are being violated is one question in an ongoing legal dispute, cf. ZackZack[48]]

7. As a result, fewer women will hold well-paid positions, which will lead to data that show that women do not work in these jobs, and to texts that describe this world, in which "a good doctor" is male. Data and texts are objective and representative.

8. Go to step 1.

Even from this highly simplified sequence alone, it is clear that not only algorithms or data are to blame. It is also obvious that data protection law cannot be called upon to mitigate all of the intermediate nor, *a fortiori*, the cumulative effects. Data are generated and collected in socio-technical systems, and algorithms operate on modelling choices and feature and category definitions that are in themselves value-laden. Decisions are not only made by companies or algorithms, but also by the affected individuals themselves, which exacerbates structural discrimination. In addition, algorithms are the back-end of systems with user interfaces and application-specific additions that were often added with good intentions but without an understanding of how they affect the system outcome as a whole.

Examples of assumptions and choices (which may further drive the feedback loops) are discussed in the analyses of public employment services algorithmic systems of several countries by *Allhutter* et al. and by *Jędrzej*, *Sztandar-Sztander-*

---

[47] *Allhutter*/*Mager*/*Cech*/*Fischer*/*Grill*, Der AMS Algorithmus – Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS), http://epub.oeaw.ac.at/?arp=0x003bfdf3/, Wien 2020.

[48] *ZackZack*, Diskriminierung und fehlende Gesetzeslage: Darum ist der AMS-Algorithmus gefährlich, https://zackzack.at/2021/05/01/diskriminierung-und-fehlende-gesetzeslage-darum-ist-der-ams-algorithmus-gefaehrlich/, 2021.

ska and *Szymielewicz*.[49] An example of assumptions is that the reasons for being unemployed lie principally with the job-seeker themselves (as reflected by the detailed way in which the individual job-seeker is modelled in the prediction model and the scarce representation of labour market factors).

An example of design choices is related to human oversight: While job counsellors can override the system's risk classification of an individual, the system requires an extra justification to be entered in such cases and thereby, especially given counsellors' time constraints, nudges them towards accepting the proposal. In a wide range of contexts, people tend to favour suggestions from automated decision-making systems and ignore contradictory information even if it is correct ("automation bias"). In the Polish public employment service system, case workers accepted the system's suggestion in 99.4 % of cases.[50] Thus, not all forms of human oversight are effective safeguards. The AI Act Proposal stipulates, in Article 14 (4) (b), that design should make users aware of automation bias, but it is unclear whether such awareness would suffice to avoid it.

The complexity of these effects is well-known in an active and growing community of researchers and practitioners, and many approaches to mitigating such discriminatory effects have been and are being developed. Still, the challenges remain, in particular since every real-life system rests on interconnected global and local design decisions, allocation decisions, and representation decisions both by human and machine actors. In addition, the very concepts of what constitutes discrimination are evolving. An overview of current approaches and open questions with a special focus on connections between informatics and legal concerns can be found in *Berendt*.[51]

To what extent does the AI Act Proposal address such complex effects? To answer this question, we need to understand how it claims to counter risks.

V. The AI Act Proposal's claim of proportionality

According to the AI Act Proposal's Explanatory Memorandum, "[t]he proposal […] is proportionate and necessary to achieve its objectives, since it follows a risk-based approach" (p. 7). This statement involves two claims.

First, the proposal weighs infringements on the rights of AI providers against risks and "imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety" (ibid., p. 7). This first claim appears validated by the construction of the law.

---

[49] *Allhutter* et al. (Fn. 47); *Jędrzej/Sztandar-Sztanderska/Szymielewicz*, Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making, https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf, 2015.

[50] *Jędrzej/Sztandar-Sztanderska/Szymielewicz* (Fn. 49).

[51] *Berendt* (Fn. 10).

Second, these "regulatory burdens" on AI providers must at the same time act as "measures" (as the GDPR would call them) for another important class of stakeholders, namely those affected by the algorithmic system. To be proportional, these measures must be suitable, necessary, and proportional in the narrow sense to achieve the goal of protecting these stakeholders' fundamental rights, such as the right to not be discriminated against.

But are they suitable?[52]

Title II's prohibitions of certain AI systems is, under the assumption that enforcement is possible, trivially suitable: a product that does not exist cannot harm anyone.

More interesting are the provisions concerning "high-risk" AI systems in Title III. These systems are allowed. The measures proposed comprise[53], in particular, transparency of various kinds, from and to various actors[54] including documentation and traceability as well as "information […] in relation to possible risks to fundamental rights and discrimination" (Recital 47). They also comprise high-quality data, human oversight, accuracy, robustness, and safety. Procedurally/structurally, a risk management system is required (Article 9).

Transparency is also the central requirement for "certain AI systems" that pose manipulation risks and that are regulated (less stringently than the so-called "high-risk" systems) in Title IV.

No doubt transparency is relevant and useful. This principle has been regarded as a cornerstone of democracy and its checks and balances for a long time, and it is nicely summarized as "sunlight is the best disinfectant".[55]

But is this enough?

VI. Limitations of transparency as a measure against discrimination

All of the examples in Section IV have been well-publicised, and phenomena like the scarce representation of women in high-profile jobs and the gender pay gap are common knowledge and the subject of many openly available datasets, many of which have high data quality and security. Arguably, there is human oversight in the labour sector, for example through employee representation and trade unions. So it seems that discrimination can also thrive under the conditions required by the AI Act Proposal.

An additional problem is that transparency, e. g. in the form of an explanation, "is probably not the remedy you are looking for" – especially when transparency is

---

[52] Since we will argue that the answer to this question is "no", we will not analyse necessity and proportionality *strictu sensu*. These questions can be the subject of future work.

[53] The wording in this paragraph is a minor re-organisation of the terms in the Explanatory Memorandum, pp. 7, 13.

[54] *Veale/Zuiderveen Borgesius* (Fn. 4), 104 ff.

[55] Modified from *Brandeis*, What publicity can do, Harper's Weekly, Dec 20, 1913, 10–13.

ensured by an explanation given ex post and the harm has already been done and is irreversible.[56] *Edwards* and *Veale* have called the overly narrow focus on explanations as an answer to data-processing challenges in the GDPR a "transparency fallacy"; the point of the present section is to argue towards the same conclusion for algorithmic systems and AI.

Power. From a sociological perspective, it can be observed that the knowledge that (and even how) discrimination has happened does not per se change existing power relations or structures.[57] The quest for a commonly acceptable fair solution requires more than just transparency and knowledge, but also power checks through rules of discourse, legal safeguards, and political deliberation in (to the extent possible) domination-free spaces of discourse.

Feedback loops. Bias and discrimination have self-reinforcing dynamics that are well-known in sociology ("cumulative causation"),[58] but that can take on unknown speed, opacity and effectiveness in technologically-enhanced decision making – the so-called "pernicious" or "runaway" feedback loops.[59] Allocative and representational *harms* are constitutive of such loops.

Feedback loops can be a challenge in systems that are fairly simple in the sense of mainly relying on one algorithm.[60]

Architecture. The challenge becomes larger when it is unclear which algorithms have which effects, in a time of IT infrastructures in which even entities perceived as one system, such as an Internet platform, are increasingly composed of huge complex and dynamic web of components.[61] The growing intractability of technical 'decisions' resulting from modern IT infrastructures can be said to extend developments in managerial decisions: as *Bagenstos* has argued, "boundaryless workplaces" in which peer decisions assume an increasingly important role, make it more difficult to apply anti-discrimination law that is patterned on identifying flaws in traditional decision making by superiors.[62] And the challenges become even more untraceable when the discriminatory effects arise in the context of larger sociotechnical systems.[63]

---

[56] *Edwards*/*Veale*, Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for, Duke Law & Technology Review, 16 (1), 2017, 18–84.

[57] E. g. *D'Ignazio*/*Klein*, Data Feminism, Cambridge (MA) 2020; *Miceli*/*Posada*/*Yang*, Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?, in: Proceedings of the ACM on Human-Computer Interaction 6 (GROUP), 2022, 1–14.

[58] *Myrdal*, An American Dilemma: The Negro Problem and Modern Democracy, New York 1944.

[59] *O'Neil* (Fn. 33); *Ensign* et al. (Fn. 33).

[60] See various examples in *O'Neil* (Fn. 33).

[61] *Gürses*/*Van Hoboken*, Privacy after the agile turn, in: Polonetsky/Tene/Selinger (Eds.), Cambridge Handbook of Consumer Privacy, Cambridge 2018, 579–601.

[62] *Bagenstos*, The structural turn and the limits of antidiscrimination law, California Law Review, 94(1), 2006, 1–47.

[63] See *McDonald*/*Barwulor*/*Mazurek*/*Schaub*/*Redmiles*, "It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online, in: 30th USENIX Security

Veale and *Zuiderveen Borgesius* have outlined further problems in the construction of the AI Act Proposal, e. g. for responsibility and accountability, that arise from an overly limited understanding of the boundaries of AI(-involving) systems and their structures and actor roles.[64] In future work, also these effects should be investigated with respect to their possible contribution to discriminatory effects.

Categories. The very categories used to describe discrimination can play an ambivalent role. On the one hand, it is commonly agreed that categories need to be observed in order to trace discrimination.[65] On the other hand, the continued use of a category can also serve to perpetuate ingroup-outgroup boundaries and bias and discrimination resulting from them.[66]

Non-unique concepts of fairness. Ethically, legally and politically, the question is whether some constellation is considered unfair at all. Who defines this, and based on which concepts of fairness?

Computational approaches are confronted with long-standing philosophical problems of the different notions of fairness and the non-commensurability between some of them. In the machine-learning literature, these problems have resurfaced in the form of different fairness metrics for data and predictions/prescriptions and the impossibility to satisfy some of them simultaneously[67] and also in the recognition that the same problem has been observed in previous attempts at formalization.[68] Political and legal scholars have pointed out that different contexts and domains are perceived as requiring different notions of fairness and thus different metrics.[69]

In addition to these considerations, practical measures are needed for involving multiple *stakeholders* (besides the developers and decision-makers) in meaningful

---

Symposium, USENIX Security, Berkeley (CA) 2021, 375–392, for an example of the interactions between loosely coupled internet platforms and the dominance of US-based ethics over the law that is applicable in users' countries. The AD reported in *McDonald* et al. is difficult to counteract also because it falls neither under European nor under German anti-discrimination law (*Pekel*, Airbnbs schwieriger Umgang mit Sexarbeiter:innen, https://netzpolitik.org/2020/diskriminierung-airbnbs-schwieriger-umgang-mit-sexarbeiterinnen/, 2020). And the AI Act Proposal would not subsume this 'trustworthiness score' under its prohibition of social scoring because the operator is a private company.

[64] *Veale*/*Zuiderveen Borgesius* (Fn. 4).

[65] Cf. the arguments about "color blindness" failing, e. g. Neville/Gallardo/Sue (Eds.), The Myth of Racial Color Blindness: Manifestations, Dynamics, and Impact, Washington, D. C. 2016.

[66] See *Bowker*/*Star*, Sorting things out: classification and its consequences, Cambridge (MA) 1999, for a comprehensive study of the effects of categorization.

[67] For a recent overview and clustering, see for example *Barocas*/*Hardt*/*Narayanan*, Fairness and Machine Learning, http://www.fairmlbook.org 2019. For their relation to legal criteria, see *Wachter*/*Mittelstadt*/*Russell*, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, Computer Law & Security Review, 2021, 41, 105567.

[68] *Hutchinson*/*Mitchell*, 50 years of test (un)fairness: Lessons for machine learning, in Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* 2019, New York City (NY) 2019, 49–58.

[69] *Binns* (Fn. 32); *Zuiderveen Borgesius* (Fn. 6).

ways, helping them express their applicable fairness notions, and resolving conflicts.

In sum, it appears that transparency, even if coupled with data quality, security, human oversight and documentation, alone is not suitable to counteract AD.

VII. Transparency is not enough: Lessons learned from data protection

In the light of these observations, we once again turn to data protection for inspiration.

Transparency is an important principle in data protection theory and data protection law. A classic argument was given by the German Constitutional Court in its 1983 Census Judgment: "A social order in which individuals can no longer *ascertain* who knows what about them and when and a legal order that makes this possible would not be compatible with the right to informational self-determination." (emphasis added).

The GDPR lists transparency as one of its fundamental principles (Article 5 (1) (a)), and it also contains requirements of documentation (through the principle of accountability, Article 5 (2)), data quality (Article 5 (1) (d)), IT security (Article 5 (1) (f)), and certain forms of human oversight ("right to obtain human intervention" under the conditions of Article 22 (3)).

However, it has long been recognized that transparency and similar requirements are not enough to protect against risks arising from the processing of personal data. Already in the paragraph immediately following the one cited, the Census Judgment adds that "the fundamental right [to informational self-determination[70]] guarantees in principle the power of individuals to *make their own decisions* as regards the disclosure and use of their personal data." – in other words, to have not only knowledge but also some extent of *control* over data processing concerning them. This has been operationalized further in the privacy protection goal of "intervenability",[71] and it has found its expression in the rights to rectification and erasure, the right to object and the rights associated with automated individual decision-making (Articles 13–22 GDPR).

Privacy and/or data protection guidelines (such as the 1980/2013 OECD Guidelines[72]) and laws (such as the GDPR in Article 5) have recognized a small

---

[70] "[…] the authority of the individual to decide himself, on the basis of the idea of self-determination, when and within what limits information about his private life should be communicated to others".

[71] *Hansen/Jensen/Rost*, Protection goals for privacy engineering, 2015 IEEE Security and Privacy Workshops, New York City (NY) 2015, 159–166.

[72] OECD, Guidelines on the protection of privacy and transborder flows of personal data. https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm (1980), https://www.oecd.org/digital/ieconomy/privacy-guidelines.htm, (2013).

but powerful set of further principles that derive from a long history of observations of risks and harms caused by the processing of personal data. *Hoepman* has provided a comparative analysis of these principles (including but not limited to the list of seven in Article 5 GDPR), for example the above as "individual participation".[73]

The principles of data minimization and purpose limitation are prime examples that illustrate that transparency alone is not sufficient to protect against abuses and unintended risks of the processing of personal data. They derive from social- and computer-science findings such as 'no data is innocent', contextual integrity as a widespread desideratum of privacy, and the fact that any re-use of data (usually in combination with re-purposing and linking previously unlinked data items) can lead to undesired re-identification and profiling of individuals.[74]

VIII. Conclusions: Legal and computer-science goals

Based on the observations and research results concerning the ways in which AD operates, and how it often operates not only (or even not at all) in isolated datasets, algorithms, or even AI systems, but in larger socio-technical systems, I consider two additions to current practices necessary. First, legal provisions (such as the AI Act) should extend beyond their current focus on transparency (and documentation, data quality, data security and human involvement in their various currently discussed forms). Second, these principles have to be transformed into system design.

1. Principles for protection against bias and discrimination

The following requirements should be turned into additional principles to be embedded into laws. As in the GDPR, these requirements should be accompanied by the requirement to deploy appropriate technical and organizational measures. This effort can draw on the lessons learned with data protection principles, but it goes beyond them because "data protection is […] less contentious than anti-discrimination. Indeed, data protection is about one particular operation (the processing of personal data), the status of which is unproblematic. Discrimination goes a step further because it does not regulate an action as such (e. g., data processing), but a legal consequence of any actions (thus, also including eventually

---

[73] *Hoepman*, Privacy design strategies, in: ICT Systems Security and Privacy Protection – 29th IFIP TC 11 International Conference, Berlin etc. 2014, 446–459.

[74] Transparency *can contribute* to control and the reduction of power asymmetries (*Naudts/Dewitte/Ausloos*, Meaningful transparency through data rights: A multidimensional analysis, in: Kosta/Leenes/Kamara [Eds.], Research Handbook on EU data protection, Cheltenham [UK] 2022) – the point is that it also may *not* do this.

data processing), which inherently entails operating a (legal) qualification of the facts."[75]

a) Bias and discrimination avoidance should be foundational. This has two aspects: *detection and prevention* (or at least mitigation).

b) Feedback loop interruption: loops should be anticipated, detected, and broken.

c) Harms recognition: the law should protect not only against risks to "health and safety" or "physical and psychological harms to individuals" (the wordings in Articles 5 and 7). These allocative and individual harms need to be supplemented by typical discrimination risks: further allocative harms and group-related representational harms.

Representational harms are difficult to operationalise, not least because one person's or group's discriminatory language is another person's or group's freedom of speech. When such language is created by a human, one can apply legal concepts such as libel, defamation, sedition and incitement to hatred and violence, or hate speech. Legal certainty should be created regarding language generated by machines (especially the responsibility and accountability for linguistic content), and state-of-the-art measures should be deployed to minimize the likelihood of such harms. Assigning responsibility and implementing measures becomes more challenging as language-processing pipelines are becoming more complex and dynamic (e. g. in architectures using pre-trained language models).

d) Fundamental rights orientation: If risk classes of AI systems are to be retained, the classification of a system should not rest on its function or operator, but its impact(s). For example, a chatbot is not per se more manipulative than, e. g., emotion recognition built into a non-chat-bot recommender system. Also, private-sector AI firms may cause "grave socioeconomic consequences [to individuals] similar to the exclusion of state-provided services".[76] The potential for harm should be assessed by a fundamental-rights impact assessment that takes into account the previous points. This principle follows the GDPR's requirement for a data protection impact assessment (Article 35) and the German Data Ethics Commission's treatment of risk classes.[77]

e) Categories awareness: The use of sensitive categories[78] should be justified in an impact assessment by an explanation of how the benefits of this use exceed its

---

[75] *Gellert* et al. (Fn. 12), p. 71.

[76] *Veale/Zuiderveen Borgesius* (Fn. 4), p. 100.

[77] See Datenethikkommission (Fn. 14). Mandatory impact assessments have also been called for by, e. g., ECNL (European Centre for Not-for-Profit Law), ECNL position statement on the EU AI Act, https://ecnl.org/news/ecnl-position-statement-eu-ai-act, 2021 and AlgorithmWatch, Draft AI Act: EU needs to live up to its own ambitions in terms of governance and enforcement, https://algorithmwatch.org/en/eu-ai-act-consultation-submission-2021/, 2021.

[78] These include, but are not limited to the GDPR's "special categories of data". See *Schwartz*, Classifying books, classifying people, https://medium.com/the-bytegeist-blog/classifying-books-classifying-people-302430282a3d, 2018, for why this need not even be limited to personal data.

risks. Such risk-benefit thinking appears to be implied by the claim to proportionality when this claim is interpreted w.r.t. the individuals affected by the AI (see Section V). In a sense, Article 10 (5) AI Act Proposal constitutes an example of Article 9 (2) (g) GDPR: Processing of special categories of personal data shall *not* be prohibited if "processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued …", with bias monitoring a processing that is necessary to minimize or avoid discrimination. This explicitness of proportionality is missing in the AI Act Proposal, and the protection of individual seems to be weaker ("appropriate safeguards" rather than the GDPR Article's "suitable and specific measures to safeguard").[79] The risk of perpetuating pernicious labelling of people through the use of categories (whether from a list of "special categories" or not) should be one of the risks considered in these analyses.

f) Metrics specificity: The concepts and metrics of fairness employed must correspond to the concepts used in the application domain; they should be derived via stakeholder-based and democratic procedures; and their choice and design should be justified in the technology impact assessments.

g) Stakeholder orientation: Stakeholders should be involved in design and processes. Involvement should include rights to transparency and participation. Who counts as a relevant stakeholder and their specific rights should be decided based on, inter alia, current debates around data protection.[80]

As *Vedder* observes, transparency obligations in data protection law have an "obvious addressee": the data subject.[81] But who would be an appropriate addressee for transparency regarding discriminatory risks if no personal data are involved and therefore no such individual can be determined? *Vedder* therefore considers transparency "not the end [but] just a beginning" and calls for a regulatory regime that enables deliberations about the possible impacts on humans. The AI Act Proposal previews public databases, such that for example consumer rights group may act as or on behalf of addressees to obtain information. However, they cannot act to intervene because "affected [individuals and] communities are provided with no mechanism for complaint or judicial redress".[82]

---

[79] The exemption of Article 10 (5) "can only be used in relation to high-risk systems, and only by those systems' providers" (*Veale/Zuiderveen Borgesius* [Fn. 4], 103). It remains to be seen whether this is useful for categories awareness or not.

[80] *Naudts* (Fn. 12).

[81] *Vedder*, Why data protection and transparency are not enough when facing social problems of machine learning in a big data context', in: Bayamlioğlu/Baraliuc/Janssens/Hildebrandt (Eds.), Being profiled: Cogitas, ergo sum. 10 Years of Profiling the European Citizen, Amsterdam 2018, 42–45. The AI Act Proposal, on the other hand, focuses on transparency towards "users", thus neglecting data subjects who are not users (*Sesing/Tschech*, AGG und KI-VO-Entwurf beim Einsatz von Künstlicher Intelligenz, MMR – Zeitschrift für IT-Recht und Recht der Digitalisierung, 2022, 25, 24–30).

[82] *Veale/Zuiderveen Borgesius* (Fn. 4), 112.

2. Software design

To transform the principles into system design, I build on the idea of privacy design patterns.[83] *Hoepman* proposed to map *principles* to *design strategies*, identify *design patterns* as the conceptual implementation of a design strategy, and then choose appropriate concrete *technologies*, in his case PETs (privacy-enhancing technologies). Table 1 shows examples for four key data protection principles.

Berendt and *Preibusch* suggested applying the design-pattern idea also to discrimination avoidance.[84] Table 2 sketches the application to principles a)–c) in Section VIII.1 above. The literature is by now too large to fit a representative selection of concrete methodologies into the table cells. The reader is referred to overviews in books[85] and encyclopedias[86], and they should ideally consult the most recent surveys in this rapidly-evolving field. Further principles (such as d)–g) above) should be transformed into design strategies, design patterns and technologies in future work.

However, designers must not forget that the 'divide and conquer' approach of design patterns cannot guarantee fair systems. As the example described by *Schaar* has illustrated, too much attention to detail in a privacy impact assessment can produce a system that, even if the best privacy-enhancing technologies are used, as a whole violates core values of data protection.[87] Just like transparency is not everything, design guidelines are not everything. A holistic view needs to complement the attention to detail.

---

[83] *Hoepman* (Fn. 73); see for an extended description *Danezis*/*Domingo-Ferrer*/*Hansen*/*Hoepman*/*Le Métayer*/*Tirtea*/*Schiffner*, Privacy and Data Protection by Design – from Policy to Engineering, https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design, 2014.

[84] *Berendt*/*Preibusch*, Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass, Big Data, 5 (2), 2017, 135–152.

[85] *Barocas*/*Hardt*/*Narayanan* (Fn. 67).

[86] *Ruggieri*, Algorithmic fairness, in: Comandé (Ed.), Elgar Encyclopedia of Law and Data Science. Cheltenham (UK) 2022; *Berendt* (Fn. 10).

[87] *Schaar*, Privacy by Design, Identity in the Information Society, 2010, 3(2), 267–274.

Principle → DESIGN STRATEGY	*Design pattern*	*Privacy-enhancing technology*
Data minimisation → MINIMISE	'select before you collect', 'anonymise and use pseudonyms', 'collect anonymous data if possible'	Anonymity metrics and anonymisation techniques, anonymous communication, …
Unlinkability, Purpose limitation → SEPARATE	"No specific design patterns" (*Danezis* et al. [Fn. 83]), but can be realized via database design: partitioning, distribution, k-anonymity, …	Specific methods for these patterns, e. g. *Jiang/Clifton*[88]
Transparency → INFORM	P3P (or rather: improved versions of its basic idea)	e. g. TILT (*Grünewald/Pallas*[89])

Table 1: From legal principles to technology choice: examples from data protection (based on *Danezis* et al. [Fn. 83]; *Hoepman* [Fn. 73]; with additions).

Principle → DESIGN STRATEGY	*Design pattern*	*Privacy-enhancing technology*
Bias and discrimination prevention & detection, Harms recognition → PREVENT-D	*Prevent* bias/discrimination	*Specific methods* (cf. *Ruggieri* [Fn. 86])
→ DEMONSTRATE-ND	*Detect* bias/discrimination	–"–
Feedback loop interruption → BREAK-LOOPS	Use data-collection procedures and machine-learning algorithms that avoid pernicious feedback loops	e. g. reinforcement learning (*Ensign* et al. [Fn. 32])

Table 2: From legal principles to technology choice: examples for the avoidance of algorithmic discrimination in future AI regulation.

[88] *Jiang/Clifton*, A secure distributed framework for achieving k-anonymity. VLDB Journal, 2006, 15(4), 316–333.

[89] *Grünewald/Pallas*, TILT: A GDPR-aligned transparency information language and toolkit for practical privacy engineering, in: ACM Conference on Fairness, Accountability, and Transparency, New York City (NY) 2021, 636–646.

KI-Trainingsdaten nach dem Verordnungsentwurf für Künstliche Intelligenz

Qualität und Performanz im Zusammenspiel von Recht und Informatik

Philipp Hacker / Lauri Wessel

I. Einleitung

Künstliche Intelligenz (KI) ist beileibe kein einheitliches Phänomen, sondern ein Sammelbegriff für unterschiedliche Formen, mit mathematischen Methoden menschlichem Handeln und Problemlösen zumindest nahezukommen.[1] Für viele dieser Techniken sind Daten notwendig, um die mathematischen Modelle zunächst zu trainieren, etwa für überwachtes Lernen oder Verstärkungslernen.[2] Die Qualität der Modelle hängt dann ganz wesentlich von der Qualität der verwendeten Daten ab. Angesichts dieser technischen Zentralität der Datengrundlage überrascht es nicht, dass in der jüngeren Regulierungsdebatte auch der rechtliche Rahmen für diese Trainingsdaten in den Vordergrund gerückt ist.[3] In Art. 10 des Kommissionsvorschlags für eine Verordnung zur Regulierung von künstlicher Intelligenz (KI-VO-E)[4] ist nun erstmalig ein expliziter Rechtsrahmen formuliert worden. Ziel dieser Vorgaben ist es insbesondere, die Qualität der Datengrundlage sicherzustellen, damit auf dieser Basis hochwertige und diskriminierungsminimierende KI-Applikationen entwickelt werden können.[5]

Dieser Beitrag unternimmt es daher, sich der Qualität und Regulierung von KI-Trainingsdaten in vier Schritten zu nähern. Zunächst werden unterschiedliche

---

[1] *Russell/Norvig*, Artificial Intelligence, 3. Aufl. 2010, 1 ff.

[2] *LeCun/Bengio/Hinton*, 521 Nature 2015, 436 (436 f.); *Sutton/Barto*, Reinforcement Learning, 2. Aufl. 2018, 2.

[3] *Pasquale*, 119 Columbia Law Review (2019), 1917; *Gerberding/Wagner*, ZRP 2019, 116; *Martini*, Blackbox Algorithmus, 2019, 228 f.; *Hacker*, ZGE 12 (2020), 239; *Hacker*, 13 Law, Innovation and Technology 2021, 257, jeweils m. w. N. Der folgende Beitrag stützt sich insbesondere für die Teile II, III und V auf die beiden letztgenannten Vorarbeiten; siehe nunmehr auch *Hacker*, GRUR 2022, 1278 zum Umgang mit Trainingsdaten bei Gatekeepern im Sinne des DMA.

[4] *Europäische Kommission*, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz), COM/2021/206 final; siehe zuletzt *Council of the European Union*, Dok. 11124/22, Second Presidency compromise text v. 15.07.2022 [dieses Dokument zitiert als: KI-VO-E Juli 2022].

[5] EG 43 und 44 KI-VO-E.

Dimensionen der Datenqualität aus dem Bereich der Informatik vorgestellt (II.). Sodann wird kurz das bestehende Regulierungsgefüge umrissen (III.). Vor diesem Hintergrund kann der KI-VO-E eingeordnet und kritisch diskutiert werden (IV.). Anschließend skizziert der Beitrag grundlegende Fragen für die Regulierung von KI-Trainingsdaten und unterbreitet einige Lösungsvorschläge mit Blick auf Rechtssicherheit, Informationswettbewerb, das Verhältnis von Ergebnis- und Prozessregulierung sowie partizipatorisches Design (V.). Eine Zusammenfassung beschließt den Beitrag (VI.).

II. Datenqualität in der Informatik

In der Informatik und der Forschung zu Informationssystemen („Information Systems" (IS) Research) besteht bereits seit Jahren eine reichhaltige Literatur, die sich mit der Frage der Qualität von Daten befasst.[6] Dabei werden unterschiedliche Dimensionen der Datenqualität entfaltet, die nicht zuletzt in der ISO/IEC 25012, einem internationalen Datenqualitäts-Standard, Niederschlag gefunden haben. Klassische Dimensionen von Datenqualität sind dabei etwa:[7]

- Richtigkeit: Die Werte der jeweiligen Attribute einer Entität (zum Beispiel Name, Alter, Einkommen einer Person) entsprechen, soweit messbar, den realen Eigenschaften der Entität.[8]
- Vollständigkeit: Zu jedem Attribut liegen für alle erfassten Entitäten auch Werte vor.[9]
- Konsistenz: Die Daten sind widerspruchsfrei.[10]
- Aktualität: Die angegebenen Werte in den Daten entsprechen den gegenwärtigen Charakteristika der jeweiligen Entitäten.[11]

All diese Maße sind graduell und nicht kategorial zu verstehen: Sie können mehr oder weniger erfüllt sein.

Die Besonderheit von KI, die mithilfe von Daten trainiert wird, liegt nun darin, dass typischerweise Aussagen nicht nur über die für die Modellierung verfügbaren Daten gemacht werden sollen, sondern auch über die konkreten Anwendungsdaten, für welche die jeweiligen Zielvariablen noch nicht bekannt sind. Beim über-

---

[6] Siehe etwa *Legner/Pentek/Otto*, 21 Journal of the Association for Information Systems (2020), 735; *Heinrich/Klier*, in: Hildebrand et al. (Hrsg.), Daten- und Informationsqualität, 4. Aufl. 2018, 47; *Lee/Strong*, 29 Journal of Management Information Systems (2003), 13; *Lee* et al., 40 Information & Management (2002), 133; *Wang/Strong*, 12 Journal of Management Information Systems (1996), 5.

[7] *Heinrich/Klier* (Fn. 6), 50 ff.; *Lee* et al., 40 Information & Management 2002, 133 (134 ff.); *Wang/Strong*, 12 Journal of Management Information Systems (1996), 5 (14).

[8] *Heinrich/Klier* (Fn. 6), 55.

[9] *Heinrich/Klier* (Fn. 6), 52.

[10] *Heinrich/Klier* (Fn. 6), 58.

[11] *Heinrich/Klier* (Fn. 6), 59.

wachten Lernen etwa wird anhand eines Datensatzes, bei dem die Zielvariable (zum Beispiel Kreditwürdigkeit) bereits bekannt ist, das KI-Modell trainiert und seine Performanz überprüft.[12] Anschließend werden damit neue, für das Modell noch unbekannte Daten analysiert (zum Beispiel von Personen, die einen Kredit beantragen und deren Kreditwürdigkeit noch unbekannt ist).

Damit ist auch klar, dass es für die Eignung der Daten zur Erstellung hochwertiger KI-Applikationen nicht nur auf die Datengrundlage für die Modellierung ankommen kann, sondern diese Daten auch in ein Verhältnis zu der beabsichtigten Anwendung und den damit verbundenen Anwendungsdaten gebracht werden müssen. Insofern werden für die Datengrundlage von KI-Modellen weitere Maße diskutiert, die über die klassischen Dimensionen von Datenqualität hinausgehen. Diese sind etwa:

- Ausgewogenheit/Balance: Die Daten sollten eine gute Balance zwischen verschiedenen rechtlich geschützten Gruppen beinhalten.[13] So sollten zum Beispiel weiße Männer nicht überrepräsentiert sein.
- Repräsentativität: Ferner sollten die Daten, die für die Modellierung verwendet werden, diejenigen Typen von Personen oder Gegenständen abdecken, die auch in der jeweiligen Anwendung als Analyseobjekt erwartet werden können.[14] So wäre etwa ein Datensatz, der nach klassischen Maßstäben hochqualitative Daten für die Kreditwürdigkeit von Studierenden bereithält, womöglich ungeeignet, um auf dieser Grundlage mit KI-Methoden die Kreditwürdigkeit von Pensionären einzuschätzen.

Inwiefern Anwendungen aus dem Bereich der KI auch klassische Datenqualitätsmaße verändern und wie diese mit neuen Dimensionen von Datenqualität interagieren, ist gegenwärtig Gegenstand der Forschung. Aus regulatorischer Perspektive lässt sich jedenfalls festhalten, dass die Erfüllung dieser Qualitätsstandards einerseits wünschenswert ist, um qualitativ hochwertige KI zu ermöglichen. In welchen Bereichen und zu welchen Zwecken derartige KI eingesetzt werden kann und darf, ist damit naturgemäß nicht entschieden. Andererseits ist die Sammlung und Aufbereitung von Daten zu KI-Trainingszwecken aufwändig und damit ressourcenintensiv, insbesondere wenn die genannten Qualitätsstandards überprüft und in hohem Maße eingehalten werden sollen.[15] Zudem ist für die Abnehmer von trainierten KI-Modellen häufig nicht leicht ersichtlich, welchen Qualitätsstan-

---

[12] *LeCun/Bengio/Hinton*, 521 Nature 2015, 436 (436 f.); *Goodfellow/Bengio/Courville*, Deep Learning, 2016, 79 ff.

[13] *Yu/Chakraborty/Menzies*, Fair Balance: Mitigating Machine Learning Bias Against Multiple Protected Attributes With Data Balancing, Working Paper 2021, arXiv preprint arXiv:2107.08310; siehe auch *Europäische Kommission*, On Artificial Intelligence – A European approach to excellence and trust, White Paper, COM(2020) 65 final, 19.

[14] *Hand*, 21 Statistical Science 2006, 1 (8).

[15] Vgl. *Chu* et al., Data cleaning: Overview and emerging challenges, Proceedings of the 2016 International Conference on Management of Data, 2016, 2201.

dards die Datengrundlage genügte.[16] Dies bedingt zugleich, dass die marktförmigen Anreize für die Ersteller der KI-Datengrundlage womöglich nicht in allen Fällen hoch genug sind, um die rigorose Einhaltung von Qualitätsstandards zu garantieren. Dies wiederum wirft die Frage auf, inwiefern diese Qualitätsstandards regulatorisch vorgegeben werden können. Dem wird im Folgenden nachgegangen.

III. Gegenwärtige rechtliche Vorgaben

Der bestehende europäische Rechtsrahmen zur Gewährleistung von Qualität hinsichtlich der Datengrundlage von KI-Applikationen kann im Rahmen dieses Beitrags nur kurz skizziert werden.[17]

1. Klassische Dimensionen von Datenqualität

Zu nennen ist hier insbesondere der Grundsatz der Datenrichtigkeit in Art. 5 Abs. 1 lit. d DS-GVO, der jedenfalls nach seinem Wortlaut jedoch nur die objektive Richtigkeit und hinreichende Aktualität von Daten verlangt, andere Dimensionen der Datenqualität dagegen außen vor lässt.[18] Bei Scoring-Verfahren verlangt zudem § 31 Abs. 1 Nr. 3 BDSG, dass die Berechnung eines Scores nicht allein auf Anschriftendaten fußen darf; es spricht allerdings viel dafür, dass diese spezifische Regelung dem Anwendungsvorrang der DS-GVO zuwiderläuft.[19] Voraussetzung für die Anwendbarkeit sowohl von DS-GVO als auch des BDSG ist allerdings bekanntermaßen, dass es sich bei den für die Modellierung verwendeten Daten um personenbezogene handelt. Dies wiederum ist insbesondere dann zumindest zweifelhaft, wenn identifizierende Informationen nachhaltig mithilfe von starken Anonymisierungsverfahren entfernt wurden.[20] So ist es durchaus nicht unüblich, identifizierende Informationen vor der Verarbeitung zu Zwecken des KI-Trainings zu eliminieren;[21] ob dann noch personenbezogene Daten nach Maßgabe des Breyer-Urteils des EuGH vorliegen,[22] dürfte sich insbesondere danach bemessen, ob eine möglicherweise technisch durchführbar, aber rechtlich verbotene Re-Identifizierung für die nach dem EuGH relevante Wahrscheinlichkeit der Herstellung

---

[16] Siehe etwa *Hacker*, ZGE 12 (2020), 239 (260 f.).

[17] Ausführlicher etwa *Stevens*, Datenqualität bei algorithmischen Entscheidungen, INFORMATIK 2019, 367; *Hacker*, ZGE 12 (2020), 239 (245 ff.); *Zech*, NJW 2022, 502.

[18] Siehe dazu etwa *Herbst*, in: Kühling/Buchner, DS-GVO/BDSG, 3. Aufl. 2020, Art. 5 Rn. 60 f.; sowie den Beitrag von *Hornung* in diesem Band.

[19] *Buchner*, in: Kühling/Buchner, DS-GVO/BDSG, 3. Aufl. 2020, § 31 BDSG Rn. 4 f.; *Moos/Rothkegel*, ZD 2016, 561 (567 f.); aA *Taeger*, ZRP 2016, 72 (74).

[20] EG 26 DS-GVO.

[21] *Oostveen*, 6 International Data Privacy Law 2016, 299, 307.

[22] EuGH, Urt. v. 19.10.2016 – Rs. C-582/14 (Breyer).

des Bezugs zu einer Person berücksichtigt werden soll.[23] Der EuGH verneint die Relevanz illegaler Re-Identifizierung,[24] sodass der Anwendbarkeit der DS-GVO insoweit signifikante Grenzen gezogen sind. Insbesondere bei synthetischen Daten fehlt von vornherein jeglicher Personenbezug. Daher ist die Tragweite der DS-GVO und auch des BDSG im hier untersuchten Bereich deutlich limitiert.

2. KI-spezifische Dimensionen von Datenqualität

Hinsichtlich der spezifischen Maße für die KI-Datengrundlage ließen sich möglicherweise das AGG oder auch der Grundsatz der Datenverarbeitung nach Treu und Glauben, Art. 5 Abs. 1 lit. a DS-GVO (*fair processing*), zur Vermeidung einer Unterrepräsentierung geschützter Gruppen mobilisieren. Bei der DS-GVO stellen sich jedoch wiederum die soeben diskutierten Probleme der Anwendbarkeit. Das AGG hingegen ist als spezifisches Instrument des Antidiskriminierungsrechts zunächst darauf ausgerichtet, bestimmte Ergebnisse zu sanktionieren, die aufgrund einer unmittelbaren oder mittelbaren Anknüpfung an verpönte Merkmale zustandekommen. Bei aller Schwierigkeit im Detail sind daher jedenfalls grundsätzlich KI-Anwendungen, die eine bestimmte nach dem AGG geschützte Gruppe benachteiligen, erfasst, sofern die jeweilige Applikation in den sachlichen Anwendungsbereich nach § 2 Abs. 1 AGG fällt.[25] Dies betrifft etwa die Nutzung von KI zu Zwecken der Personalauswahl, der Auswahl von Studierenden oder von Mietern.

Noch nicht abschließend geklärt hingegen ist, inwiefern das AGG Anwendung finden kann auf Tätigkeiten, die im Vorfeld eines derartigen Einsatzes von KI liegen – etwa auf die bloße Zusammenstellung von Trainingsdaten.[26] Der EuGH hat etwa klargestellt, dass Äußerungen eines Arbeitgebers, die einem Einstellungsverfahren vorgelagert sind, ebenfalls dem Schutz des AGG unterfallen, wenn sie einen hinreichenden Bezug zu einem Einstellungsverfahren aufweisen.[27] In ähnlicher Weise ließe sich argumentieren, dass die Zusammenstellung von Trainingsdaten vom AGG erfasst ist, wenn ein klarer, nicht bloß hypothetischer Bezug zu einer Erstellung eines KI-Modells besteht, das in einem nach Maßgabe des § 2 Abs. 1 AGG relevanten Kontext eingesetzt werden soll. Wie genau eine relevante Benachteiligung in einem solchen Fall nachgewiesen und gegebenenfalls gerechtfertigt werden kann, bedürfte näherer Untersuchung.

---

[23] Dazu etwa *Bergt*, ZD 2015, 365 (370); *Karg*, in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, 2019, Art. 4 Nr. 1 DS-GVO Rn. 64; *Hacker*, Datenprivatrecht, 2020, 108 ff.

[24] EuGH, Urt. v. 19.10.2016 – Rs. C-582/14 (Breyer) – Rn. 46.

[25] Dazu ausführlich der Beitrag von *Grünberger/Müller* in diesem Band m. w. N.; siehe auch *Wachter/Mittelstadt/Russell*, 41 Computer Law & Security Review (2021), 105567; *Hacker*, 55 Common Market Law Review (2018), 1143 (1154 ff.); *Thüsing*, in: MüKo BGB, 9. Aufl. 2021, § 3 AGG Rn. 34.

[26] Dazu auch *Hacker*, ZGE 12 (2020), 239 (251 ff.).

[27] EuGH, Urt. v. 23.4.2020 – Rs. C-507/18 (Associazione Avvocatura per i diritti LGBTI) – Rn. 39, 58; Urt. v. 10.7.2008 – Rs. C-54/07 (Feryn) – Rn. 28, 34.

Umgekehrt stellt sich ferner das Problem, dass die Herstellung von Ausgewogenheit in einem Trainingsdatensatz einhergehen kann mit der Verarbeitung von sensiblen personenbezogenen Daten nach Art. 9 Abs. 1 DS-GVO.[28] Darunter fallen etwa Angaben über die ethnische Herkunft, die Religion oder die sexuelle Orientierung der jeweiligen Personen. Wenn Entwickler derartige Angaben verarbeiten, um einen Datensatz diverser zu gestalten, verstoßen sie daher möglicherweise gegen Art. 9 Abs. 1 DS-GVO. Zwar ließe sich argumentieren, dass eine derartige Reduktion von Diskriminierungspotenzial „aus Gründen eines erheblichen öffentlichen Interesses erforderlich" ist und daher nach Art. 9 Abs. 2 lit. g DS-GVO erlaubt sein könnte (sofern dessen andere Tatbestandsmerkmale erfüllt sind). Allerdings handelt es sich hierbei lediglich um eine Öffnungsklausel, nicht um eine bereits für sich geltende Ausnahme von Art. 9 Abs. 1 DS-GVO.[29] Die hierbei bestehende erhebliche Rechtsunsicherheit und Haftungsgefahr schwächen jedoch die Anreize, diskriminierungssensitive Datensätze zu erstellen.

Konkrete Normen zur Gewährleistung von Repräsentativität bei der Datengrundlage von KI-Applikationen finden sich gegenwärtig nicht. Anreize dafür ließen sich allenfalls aus dem allgemeinen Haftungsrecht herleiten. Hier ist jedoch vieles noch ungeklärt, insbesondere was auch die Anwendbarkeit und Voraussetzungen einer Haftung für mangelhafte Trainingsdaten nach dem Produkthaftungsrecht anbelangt.[30] Darauf wird noch zurückzukommen sein. Insgesamt zeigt sich damit, dass auch hinsichtlich KI-spezifischer Dimensionen von Datenqualität der bestehende Rechtsrahmen die im ersten Abschnitt genannten Qualitätsmaße nur unzureichend berücksichtigt.

IV. Art. 10 KI-VO-E

Diese Leerstelle erklärt die umfangreiche Regelung, welche die Kommission in ihrem Verordnungsvorschlag der Datengrundlage von KI-Modellen widmet. Art. 10 KI-VO-E stellt eine Reihe von Rahmenbedingungen auf, die erfüllt sein müssen, wenn Daten zur Erstellung von Hochrisiko-Modellen genutzt werden. Dabei unterscheidet der Vorschlag begrifflich zwischen Trainings-, Validierungs-

[28] Siehe etwa *Dwork* et al., Fairness through awareness, Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 2012, 214; *Veale/Binns*, 4 Big Data & Society (2017), 205395171774353; *Andrus* et al., What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 249.

[29] *Petri*, in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, 2019, Art. 9 DS-GVO Rn. 71.

[30] Dazu nun *Zech*, NJW 2022, 502; siehe auch *Schuhmacher/Fatalin*, CR 2019, 200 (204); *Hacker*, ZGE 12 (2020), 239 (250).

Klassische Maße:	Rechtsrahmen:	Spezifische Maße für KI-Trainingsdaten:	Rechtsrahmen:
1. Richtigkeit	Art. 10 III KI-VO-E	1. Balance	Art. 10 III KI-VO-E?
2. Vollständigkeit	Art. 10 III KI-VO-E	2. Repräsentativität	Art. 10 III/IV KI-VO-E
3. Konsistenz	(–)		
4. Aktualität	Art. 5 I lit. d DS-GVO		

Figur 1: Übersicht über verschiedene Maße für Datenqualität und ihre rechtliche Verankerung unter Berücksichtigung des KI-VO-E.

und Testdaten;[31] sofern nicht weiter unterschieden, sind in diesem Beitrag mit dem Begriff der „Trainingsdaten" alle drei Datentypen gemeint. Denn die zentralen Vorgaben des Vorschlags gelten für sie alle gemeinsam.

1. Überblick über die wesentlichen Neuerungen

Art. 10 KI-VO-E stellt einen Governance-Rahmen für KI-Trainingsdaten auf, der freilich wie alle Vorschriften von Titel III des KI-VO-E nur für KI-Systeme gilt, die als Hochrisikosysteme eingestuft werden. Darunter fallen sog. „embedded systems", die in Produkte integriert werden, für die produktsicherheitsrechtliche Zulassungsverfahren gelten (etwa Medizinprodukte),[32] sowie „standalone systems", die aufgrund ihrer Verwendung ein besonderes Risiko darstellen sollen.[33] Dies betrifft einerseits primär öffentlich-rechtlich geprägte Gebiete wie Rechtsdurchsetzung und Grenzkontrollen, andererseits aber auch privatrechtlich konturierte Bereiche wie die Personalauswahl oder das Kreditscoring.

Die Governance-Vorschriften sollen dabei nach Art. 10 Abs. 2 KI-VO-E den gesamten Lebenszyklus der Trainingsdaten umfassen, von der Konzeption über die Sammlung und Aufbereitung bis hin zu ihrer Evaluation. Das Kernstück der Regelung findet sich dann in den Abs. 3 und 4 (siehe auch Figur 1).

a) Klassische Dimensionen von Datenqualität

In den geplanten Vorgaben finden sich zwei klassische Maße für Datenqualität: Trainingsdaten müssen die Kriterien der Richtigkeit und Vollständigkeit erfüllen (Art. 10 Abs. 3 S. 1 KI-VO-E). Hinsichtlich der Richtigkeit stellt die Vorschrift damit eine *lex specialis* zum Grundsatz der Datenrichtigkeit aus Art. 5 Abs. 1 lit. d DS-GVO dar (soweit diese überhaupt anwendbar ist). Der Begriff der Vollständig-

---

[31] Trainingsdaten werden nach Art. 3 Nr. 29 KI-VO-E zur Anpassung der lernbaren Parameter genutzt, Validierungsdaten nach Art. 3 Nr. 30 KI-VO-E zu einer Bewertung des Lernprozesses und zur Kalibrierung der Hyperparameter. Testdaten hingegen werden während des Trainingsvorgangs nicht genutzt und dienen der Bewertung der Performanz des Modells vor Inverkehrbringen, Art. 3 Nr. 31 KI-VO-E.

[32] Art. 6 Abs. 1 KI-VO-E.

[33] Art. 6 Abs. 2 KI-VO-E i. V. m. Anhang III.

keit kommt neu hinzu. Hinsichtlich der Aktualität verbleibt es jedoch merkwürdigerweise bei der Regelung in der DS-GVO. Hier wäre Art. 10 Abs. 3 KI-VO-E noch nachzuschärfen, sofern man die dort aufgeführte Relevanz der Trainingsdaten nicht mit dem Begriff der Aktualität auflädt.[34] Dieser Begriff ist allerdings allgemein problematisch. Relevanz wird von den IS-Forschenden *Lee* und *Strong* z. B. definiert als „the extent to which data are applicable and helpful for the task at hand."[35] Dies lässt sich für einzelne Trainingsdatenpunkte gerade bei komplexen Modellen aufgrund von deren Opazität jedoch nur schwer beantworten und jedenfalls erst dann, wenn das Training bereits durchgeführt wurde.[36] Ziel der Datenanalyse ist es ja häufig, auch solche Korrelationen zu erfassen, die für den menschlichen Beobachter bei unbefangener Betrachtung gerade nicht offensichtlich relevant erscheinen. Damit bleibt die Operationalisierung des Relevanzbegriffs unklar. Das Kriterium der Konsistenz ist weiterhin weder in der DS-GVO noch in dem Verordnungsentwurf zu finden.

b) KI-spezifische Dimensionen von Datenqualität

Als echte Neuerungen kommen nunmehr Regelungen hinzu, welche auch KI-spezifische Maße für Trainingsdatenqualität in den Blick nehmen. Art. 10 Abs. 3 S. 1 KI-VO-E ordnet zunächst die Repräsentativität der Trainingsdaten an. Dies wird ergänzt durch die Regelung in Art. 10 Abs. 4 KI-VO-E, wonach die Trainingsdaten die typischen geographischen, verhaltensbezogenen oder funktionalen Rahmenbedingungen abbilden müssen, die bei bestimmungsgemäßer Verwendung der konkreten Hochrisiko-KI-Applikation zu erwarten sind. Ein System, das Kandidaten bei Personalauswahl vorselektiert, darf mithin nicht lediglich auf Daten von US-amerikanischen Bewerbern trainiert werden, wenn es in Deutschland eingesetzt werden soll. Art. 42 Abs. 1 KI-VO-E enthält insoweit eine (tautologisch anmutende) Konformitätsvermutung, sofern die Trainingsdaten entsprechend den geographischen, verhaltensbezogenen und funktionalen Rahmenbedingungen ausgewählt wurden.

Einen möglichen Anknüpfungspunkt für die Ausgewogenheit des Datensatzes zwischen nichtdiskriminierungsrechtlich geschützten Gruppen bietet Art. 10 Abs. 3 S. 2 KI-VO-E. Danach müssen die Trainingsdaten geeignete statistische Merkmale besitzen bezüglich der Personen oder Personengruppen, auf die das System bestimmungsgemäß angewandt werden soll. Diese Personengruppen

---

[34] Siehe etwa *Wang/Strong*, 12 Journal of Management Information Systems (1996), 5 (9): „The data must be relevant to the consumer. For example, […] timely for use by the data consumer in the decision-making process." Siehe auch ebd., 16.

[35] *Lee/Strong*, 29 Journal of Management Information Systems (2003), 13 (18).

[36] Siehe zur parallelen Problematik der Erheblichkeit von Daten in § 31 Abs. 1 Nr. 2 BDSG: *Sachverständigenrat für Verbraucherfragen*, Verbrauchergerechtes Scoring, Gutachten, 2018, 131 f., 144; *Domurath/Neubeck*, Verbraucherscoring aus Sicht des Datenschutzrechts, Working Paper für den Sachverständigenrat für Verbraucherfragen, 2018, 24; *Gerberding/Wagner*, ZRP 2019, 116 (118).

erschöpfen sich zwar nicht in den nichtdiskriminierungsrechtlich geschützten; sofern diese nach dem AGG geschützten Gruppen sich in statistischer Hinsicht unterscheiden, sind sie aber in jedem Fall zu berücksichtigen und adäquat abzubilden. Soll also etwa ein Erkennungssystem für Hautkrebs in einer ethnisch diversen Bevölkerungsgruppe eingesetzt werden, so darf es nicht lediglich auf Daten trainiert werden, bei denen die Hautfarbe der Mehrheit der Bevölkerungsgruppe entspricht, sondern diese müssen unterschiedliche Schattierungen von Hautfarbe beinhalten.[37] Dabei können jedoch auch weitere Personengruppen eine Rolle spielen, etwa wenn signifikante Unterschiede an den sozioökonomischen Status gekoppelt sind. Auch die hier in Rede stehende Regelung knüpft jedoch an die bestimmungsgemäße Verwendung an, sodass dadurch letztlich keine allgemeine Balance zwischen allen denkbaren geschützten Gruppen angeordnet, sondern vielmehr die gruppenspezifische Repräsentativität des Datensatzes noch einmal gesondert betont wird. Dies ist insofern nachvollziehbar, als eine generelle und vollständige Ausgewogenheit des Datensatzes zwischen allen denkbaren geschützten Gruppen regelmäßig an der mangelnden Verfügbarkeit von Daten aus allen Subgruppen scheitern dürfte.

Dies deutet auf ein allgemeines Problem hin: Insgesamt erscheint sowohl hinsichtlich der klassischen als auch der KI-spezifischen Qualitätsmaße problematisch, dass diese nach dem Wortlaut des Art. 10 KI-VO-E vollständig erfüllt werden müssen, um eine Verletzung der Vorschriften zu vermeiden.[38] Bei großen Datensätzen wird es jedoch kaum einmal vermeidbar sein, dass sich etwa einzelne Fehler in den Datensätzen finden.[39] Diese müssen nicht einmal eine signifikante Auswirkung auf das Ergebnis haben, würden jedoch nach dem bisherigen Entwurf die Fehlerfreiheit der Trainingsdaten vollständig infrage stellen. Dies steht quer zu der Erkenntnis, dass in der Informatik auch eine graduelle Erfüllung der Qualitätsmaße möglich ist.

c) Re-balancierung von Datensätzen

Eine gruppenspezifische Repräsentativität von Trainingsdatensätzen lässt sich jedoch häufig, wie gesehen, nur dadurch erzielen, dass sensitive personenbezogene Daten (z. B. zur ethnischen Zugehörigkeit, Religion oder sexuellen Orientierung) verarbeitet werden. Hier sieht Art. 10 Abs. 5 KI-VO-E nun einen spezifischen, auf die Öffnungsklausel des Art. 9 Abs. 2 lit. g DS-GVO gestützten Erlaubnistatbestand

---

[37] *Goyal* et al., Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities, 127 Computers in Biology and Medicine (2020), 104065, 7.

[38] *Smuha* et al., How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act, Working Paper 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991, 34.

[39] *Floridi*, 34 Philosophy of Technology (2021), 215 (219); *Northcutt* et al., Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks, Working Paper 2021, arXiv preprint arXiv:2103.14749.

vor, der eine derartige Verarbeitung unter bestimmten Rahmenbedingungen gestattet. Demnach dürfen sensitive personenbezogene Daten in Trainingsdatensätzen verarbeitet werden, sofern dies zur „Beobachtung, Erkennung und Korrektur von Verzerrungen unbedingt erforderlich" ist, wobei Vorkehrungen für den Schutz von Grundrechten der betroffenen Personen getroffen werden müssen. Darunter zählen etwa, wo möglich, Anonymisierung, sonst Pseudonymisierung, Verschlüsselung und Maßnahmen zur Verhinderung einer Weitergabe, Art. 10 Abs. 5 aE KI-VO-E.

Diese an sich sinnvolle Regelung ist allerdings auf Hochrisiko-Systeme beschränkt, sodass die Probleme der datenschutzrechtlichen Erlaubnis der Rebalancierung von Datensätzen für alle anderen KI-Systeme fortbestehen. Ferner ist das Verhältnis zum bestehenden Nichtdiskriminierungsrecht unterspezifiziert, wie gleich noch im Einzelnen zu erörtern ist.

2. Bewertung

Die Bewertung von Art. 10 Abs. 5 KI-VO-E fällt durchaus zwiespältig aus.

a) Vorzüge

Die Regelungen zur Daten-Governance stellen einen Schritt in die richtige Richtung dar. Drei Aspekte sind hier besonders hervorzuheben. Erstens wird durch die Regelungen die Anwendbarkeit von Qualitätsvorschriften auf Trainingsdaten geklärt, unabhängig von der umstrittenen Anwendbarkeit der DS-GVO oder einer etwaigen Ausdehnung des Schutzes durch das AGG. Zweitens werden die Maße für Datenqualität gegenüber der Regelung im DS-GVO-Grundsatz der Datenrichtigkeit umfassender genannt, indem insbesondere die Vollständigkeit, aber auch die Repräsentativität sowie die verwendungsspezifische Ausgewogenheit zwischen geschützten Gruppen explizit aufgenommen werden. Drittens wird durch Art. 10 Abs. 5 KI-VO-E eine neue Ausnahme zu Art. 9 Abs. 1 DS-GVO hinzugefügt, die in dem eng umrissenen Kontext der Datenverarbeitung zu Zwecken der Diskriminierungsvermeidung auch die Verarbeitung von sensitiven Daten zulässt. Auch hierdurch wird eine Lücke im bestehenden Datenschutzrecht geschlossen.

b) Kritik

Auf einzelne Schwierigkeiten bei der Auslegung von Art. 10 KI-VO-E oder markante Leerstellen wurde bereits in der überblicksartigen Vorstellung eingegangen. Aus einer übergeordneten Perspektive zeigen sich insbesondere drei größere Probleme hinsichtlich des vorgeschlagenen Regimes der Daten-Governance.

Erstens ist eine mangelnde Koordination mit den Nichtdiskriminierungsregeln festzustellen. Denn Art. 10 Abs. 2 lit. f KI-VO-E spricht zwar davon, dass eine Untersuchung im Hinblick auf mögliche Verzerrungen erfolgen soll, und auch

der geplante Erlaubnistatbestand in Art. 10 Abs. 5 KI-VO-E knüpft an „Verzerrungen" an. Ob damit jedoch echte (unmittelbare oder mittelbare, gerechtfertigte oder nicht gerechtfertigte) Diskriminierungen gemeint sind, bleibt im Dunkeln.[40] Vielmehr legt die Wortwahl, die sich bewusst von der Terminologie des Nichtdiskriminierungsrechts unterscheidet, nahe, dass der Begriff der Verzerrung nicht notwendig deckungsgleich mit dem der Diskriminierung sein muss. Dies hätte jedoch zur Folge, dass entweder nach dem AGG gerechtfertigte Diskriminierungen dennoch einen Handlungsbedarf nach Art. 10 Abs. 2 KI-VO-E auslösen könnten, oder gar ungerechtfertigte Diskriminierungen nicht unter die Untersuchungspflicht und die Korrekturmöglichkeit von Art. 10 Abs. 2 und 5 KI-VO-E fielen. Letzteres ist nur schwer vorstellbar.

Für die erste Möglichkeit, einen weiteren Verzerrungsbegriff als den der ungerechtfertigten Diskriminierung, spricht in gewisser Weise, dass es das Regime der Daten-Governance von schwierigen Abwägungsfragen hinsichtlich der Rechtfertigung von Benachteiligungen entlasten würde. Relevant wäre dann möglicherweise lediglich eine Ungleichverteilung in den Daten, die erwartbarerweise zu einer AGG-rechtlich relevanten statistischen Ungleichverteilung der KI-Ergebnisse in der bestimmungsgemäßen Anwendung führen würde, unabhängig davon, ob diese Benachteiligung rechtfertigbar ist. Derartige Ungleichverteilungen in den Daten müssten dann geprüft und dokumentiert werden (Art. 10 Abs. 2 KI-VO-E i. V. m. Art. 13 Abs. 3 lit. b v) KI-VO-E und Art. 11 Abs. 1, Anhang IV Abs. 2 lit. d KI-VO-E) und könnten – müssten jedoch nicht – nach Art. 10 Abs. 5 KI-VO-E bereinigt werden. Eine Korrektur*pflicht* hinsichtlich der Trainingsdaten könnte sich nach dieser Lesart etwa aus Art. 10 Abs. 3 S. 2 KI-VO-E ergeben, wenn etwa die Performanz des KI-Modells wegen einer ungleichen Repräsentation von Gruppen in den Trainingsdaten zwischen diesen Gruppen unterschiedlich ausfällt. Dies haben Forschende etwa für Gesichtserkennungssysteme nachgewiesen, die insbesondere bei Women of Color höhere Fehlerraten aufweisen.[41] Angesichts der in fast allen Bereichen vorgesehenen Selbstzertifizierung der Erfüllung der Verpflichtungen aus den Art. 7 ff. KI-VO-E[42] bleiben jedoch die im Nichtdiskriminierungsrecht generell beobachteten Durchsetzungsprobleme[43] grundsätzlich bestehen.

Ein zweiter, allgemeiner Kritikpunkt besteht darin, dass einerseits der *Prozess* des Modelltrainings durch Art. 10 KI-VO-E direkt reguliert wird, andererseits

---

[40] Vgl. *Ebers* et al., 4(4) J (2021), 589 (596).

[41] *Buolamwini/Gebru*, Conference on Fairness, Accountability and Transparency in Machine Learning (FAT*) 2018, 77.

[42] Siehe Art. 43 KI-VO-E.

[43] *Wachter/Mittelstadt/Russell*, 41 Computer Law & Security Review (2021), 105567; *Hacker*, 55 Common Market Law Review (2018), 1143 (1167 ff.); *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, 107 f.; siehe auch den Beitrag von *Grünberger/Müller* in diesem Band.

jedoch auch eine Kontrolle der KI-*Ergebnisse* durch das allgemeine Haftungsrecht und die Vorgaben zur Performanz in Art. 15 KI-VO-E installiert wird. Nach der letztgenannten Norm müssen Hochrisiko-Systeme ein angemessenes Maß an Genauigkeit (*accuracy*) erreichen.[44] Diese Ergebnisregulierung wirkt auf die Trainingsdaten zurück insofern, als mangelhafte Trainingsdaten zu suboptimalen Ergebnissen und damit einer Verletzung der Vorgaben des allgemeinen Haftungsrechts oder von Art. 15 KI-VO-E führen können. Letztlich kommt es damit zu einer Doppelung von direkter und indirekter Regulierung der Trainingsdaten. Daran knüpft sich die Frage, ob eine derartig engmaschige und für die Entwickler mit hohem Dokumentations- und Kostenaufwand einhergehende Doppelung notwendig ist. Hier ist sicherlich das letzte Wort noch nicht gesprochen. Für die Flankierung der indirekten Regulierung auch durch eine direkte Form der Regulierung spricht sicherlich, dass die Entwickler der Trainingsdaten und die Anwender des KI-Modells unterschiedliche Entitäten sein können, in welchem Fall die indirekte Regulierung Anreize für die Entwickler der Trainingsdaten zur Fehlervermeidung lediglich insoweit setzt, als diese mit einem Rückgriff durch die Anwender rechnen müssen.[45] Dies ist jedoch angesichts vertraglicher Gestaltungsmöglichkeiten sowie der nicht immer einfach zu bestimmenden Ursache von Fehlern im KI-Ergebnis durchaus nicht garantiert, sodass aus einer Anreizperspektive durchaus einiges für die direkte Regulierung spricht. Dennoch sollten Prozessregulierung und Ergebnisregulierung besser aufeinander abgestimmt werden, wie der fünfte Teil dieses Beitrags im Einzelnen zeigen wird.

Der dritte allgemeine Kritikpunkt geht dahin, dass Art. 10 Abs. 3 KI-VO-E durch eine Reihe von unscharfen Begriffen nicht unerhebliche Rechtsunsicherheit erzeugt, etwa durch die Nennung von „geeigneten statistischen Merkmale[n]“[46] oder die bereits diskutierte „Relevanz“ von Trainingsdaten. Auch der Begriff der „Verzerrung“ ist gegenwärtig, wie gezeigt, nicht klar umrissen. Dies ist grundsätzlich bei neuen Rechtsvorschriften nichts Ungewöhnliches und erzeugt eine gewisse Flexibilität in der Anwendung, die jedoch insofern fragwürdig ist, als gerade Art. 10 KI-VO-E gemeinsam mit Art. 5 KI-VO-E zu den zwei Vorschriften zählt, deren Verletzung die Maximalsanktion von bis zu 6 % des weltweiten Jahresumsatzes nach Art. 71 Abs. 3 KI-VO-E nach sich ziehen kann. Dies wirft die Frage

---

[44] Zweckdienlicher wäre es allerdings, hier allgemeiner von Performanz statt Genauigkeit zu sprechen, da die Genauigkeit (*accuracy*) nur eines von vielen möglichen Performanzmaßen darstellt, siehe *Goodfellow/Bengio/Courville*, Deep Learning, 2016, 100 f., 410 f.

[45] Denkbar wäre allenfalls noch eine direkte Haftung als Teilprodukt- oder Grundstoff-Hersteller nach dem Produkthaftungsgesetz, § 4 Abs. 1 S. 1 ProdHaftG, siehe auch *Zech*, NJW 2022, 502 (504, 507).

[46] *Smuha* et al., How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act, Working Paper 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991, 33 f.; siehe zu einem Interpretationsvorschlag von Geeignetheit im Sinne von Diskriminierungsfreiheit *Hacker*, 13 Law, Innovation and Technology (2021), 257 (298).

auf, wie den genannten Kritikpunkten begegnet und der begriffliche Gehalt des Daten-Governanceregimes geschärft werden kann. Dem wendet sich der abschließende Teil des Beitrags zu.

V. Reform- und Lösungsvorschläge

Die bisherigen Erwägungen haben eine Reihe von Problemen offenbart, welche die Notwendigkeit einer Anpassung von Art. 10 KI-VO-E im Laufe des Gesetzgebungsverfahrens bzw. einer spezifischen, technikadäquaten Interpretation nahelegen. Dies sind insbesondere Fragen des Anerkenntnis des graduellen Charakters von Qualitätsmaßen (1.), des Wettbewerbs durch die Offenlegung von Metadaten (2.), der Koordinierung der Ergebnis- mit der Prozessregulierung (3.), und schließlich der Gestaltung von Trainingsdatensätzen mit Blick auf vertrauenswürdige KI und Prozessgerechtigkeit (4.).

1. Rechtssicherheit durch risikospezifische safe harbors*: Gradueller Charakter von Qualitätsmaßen*

Wie bereits in der informatischen Übersicht erwähnt, handelt es sich bei den Qualitätsmaßen nicht um binäre Kategorisierungen, sondern um graduelle Abstufungen. Dies sollte auch die Interpretation der entsprechenden Anforderungen in der KI-VO-E anleiten. Ob ein Datensatz, der aus 100.000 Datenpunkten besteht, 50 oder 5000 Fehler enthält, ist für die Qualität ebenso bedeutsam wie die Fragen, ob es sich bei den Fehlern um kleine oder um große Abweichungen vom korrekten Wert handelt und welche Attribute betroffen sind. Alle Fehlerarten würden jedoch nach dem Wortlaut zum Verdikt der Fehlerhaftigkeit des Trainingsdatensatzes führen und könnten lediglich im Rahmen der Sanktionen unterschiedlich gewertet werden. Da jedoch gewisse Fehler, Leerstellen und Imperfektionen bei der Behandlung von großen Datensätzen unvermeidbar sind,[47] erscheint es überzeugender, in Abhängigkeit von der Risikoträchtigkeit der KI-Applikation Toleranzbereiche unterschiedlicher Größe zu definieren, innerhalb derer jedenfalls grundsätzlich davon ausgegangen werden kann, dass die Anforderungen an die Qualität der Trainingsdaten noch eingehalten wurden.[48] Zudem sollte es möglich sein, Modifikationen an den Datensätzen (etwa aus dem Bereich der *differential privacy*[49]) vorzunehmen, die auf einen Schutz der Privatsphäre abzielen, auch wenn dies zu einer höheren Fehlerquote führen sollte.[50] EG 44 der KI-VO-E

[47] Siehe oben, Fn. 39.

[48] *Floridi*, 34 Philosophy of Technology (2021), 215 (219); *Hacker*, 13 Law, Innovation and Technology (2021), 257 (299).

[49] *Dwork*/*Roth*, 9 Foundations and Trends in Theoretical Computer Science (2014), 211.

[50] So zu Recht *Ebers* et al., 4(4) J (2021), 589 (595 f.).

geht bereits in diese Richtung: Danach müssen Trainingsdatensätze nur „hinreichend relevant" etc. sein.[51] Bei KI-Applikationen, die Leib und Leben betreffen (z. B. Medizinprodukte), wären diese „erlaubten Fehlerquoten" kleiner als bei Systemen, die zwar auch der Hochrisikokategorie unterfallen, jedoch regelmäßig ein geringeres Risiko für die Betroffenen darstellen, wie etwa Systeme der Personalauswahl oder des Kreditscoring. Hierin zeigt sich einmal mehr, dass die monolithische Hochrisikokategorie des KI-VO-E, in der ganz unterschiedliche KI-Systeme und Anwendungen gebündelt werden, tendenziell unterkomplex ist und durch einen risikobasierten Interpretationsansatz ausdifferenziert werden muss.

Die Entwicklung von derartigen Toleranzbereichen ist ein genuin interdisziplinäres Unterfangen, das beispielsweise in harmonisierte Normen nach Art. 40 KI-VO-E oder in gemeinsame Spezifikationen, festgelegt in delegierten Rechtsakten auf unionaler Ebene, nach Art. 41 Abs. 1 KI-VO-E münden kann.[52] Dies hätte den Vorteil, dass prüfbare qualitative Schwellen als *safe harbors* erarbeitet werden, die mit einem hohen Maß an Vorhersehbarkeit und Rechtssicherheit für die Entwickler einhergehen.[53] Soweit die Trainingsdaten sich innerhalb der vorgesehenen Toleranzbereiche für die Qualitätsmaße bewegen, würde gemäß Art. 40 bzw. Art. 41 Abs. 3 KI-VO-E die Konformität mit den Vorgaben des Art. 10 KI-VO-E vermutet. Entscheiden sich die Entwickler hingegen, diese Toleranzbereiche zu verlassen, so agieren sie gewissermaßen auf eigenes Risiko, da dann regelmäßig eine Verletzung der entsprechenden Vorgaben in Rede stehen wird, wenn nicht besondere Umstände dem entgegenstehen.[54] Zugleich besteht eine entscheidende Herausforderung darin, die Standardisierungsabläufe transparent und für alle Stakeholder gestaltbar zu machen, da hier die großen normativen Fragen nach dem akzeptablen Risiko konkret entschieden werden.[55]

2. Wettbewerb durch Informationen über Metadaten

Auch die Offenlegung von Informationen über Trainingsdaten kann dazu beitragen, die Qualität der Datengrundlage zu gewährleisten, z. B. durch Marktdruck und Reputation. Wie die breite Diskussion der Verhaltensökonomie im Verbraucherrecht gezeigt hat,[56] sind solche Offenlegungen jedoch nur dann relevant,

---

[51] Siehe auch *Veale/Zuiderveen Borgesius*, CRi 2021, 97 Rn. 41; nach dem Wortlaut unklar war allerdings, ob sich die Modifikation „hinreichend" nur auf das Relevanzkriterium oder auch auf die übrigen Qualitätskriterien beziehen soll; siehe nunmehr aber EG 44 KI-VO-E Juli 22 (Bezug auf Relevanz und Repräsentativität) und Art. 10 Abs. 3 KI-VO-E Juli 22: „to the best extent possible, free of errors and complete".

[52] *Veale/Zuiderveen Borgesius* (Fn. 51), Rn. 50 f.

[53] Siehe bereits *Hacker*, 13 Law, Innovation and Technology (2021), 257 (299).

[54] *Veale/Zuiderveen Borgesius* (Fn. 51), Rn. 53.

[55] *Veale/Zuiderveen Borgesius* (Fn. 51), Rn. 54 ff.

[56] Siehe nur *Zamir/Teichman*, Behavioral Law and Economics, 2018; *Hacker*, Verhaltensökonomik und Normativität, 2017, 395 ff., 866 ff.; *Schmolke*, Grenzen der Selbstbindung im Privatrecht, 2014, 705 ff.

wenn sie die kognitiven Grenzen der jeweiligen Empfänger respektieren. Es sollte allerdings nicht vergessen werden, dass solche Informationen auch für KI-Experten wie Entwickler oder professionelle Betreiber von KI-Systemen nützlich sein können. Darüber hinaus können Informationen von Informationsintermediären wie Berufsverbänden, NGOs oder Forschenden gesammelt, analysiert und der Öffentlichkeit zugänglich gemacht werden.[57] Wichtig ist, dass derartige Transparenz in Art. 13 Abs. 3 lit. b v) KI-VO-E vorgesehen ist, der die Offenlegung aller „relevanter Informationen" über die Trainingsdaten vorschreibt.

Diese Offenlegungen sollten jedoch nicht auf KI-Systeme mit hohem Risiko beschränkt sein, wie in Artikel 13 KI-VO-E, sondern allgemein für Trainingsdaten gelten. Letztlich können zwei Arten von Informationen nützlich sein: eine von Experten für Experten und eine von Experten für Endnutzer. Vor allem bei der letztgenannten Art der Offenlegung werden die kognitiven Beschränkungen, die in der Verhaltensökonomie betont werden, von größter Bedeutung sein. Empfängerorientierte Gestaltungsregeln könnten in einem einfachen Trainingsdaten-Label enden, ähnlich dem Energieeffizienz-Label.[58] Mit diesen beiden Arten der Offenlegung könnte die Informationsasymmetrie bezüglich der Qualität von Trainingsdatensätzen adressiert und ein Markt für hochwertige KI-Trainingsdaten gestärkt werden.

3. Verhältnisbestimmung: Primat der Ergebnisregulierung

Ein zentrales Ziel des europäischen Ansatzes der Gestaltung von KI ist es, Anreize für effektive, performativ hochwertige KI zu entwickeln (siehe etwa EG 3 und 5 KI-VO-E).[59] Daraus lässt sich folgern, dass bei Nachweis einer hohen Ergebnisqualität möglicherweise nicht alle Anforderungen hinsichtlich des Prozesses erfüllt sein müssen. Konkret erscheint ein „Einwand hinreichender Ergebnisqualität" sinnvoll. Dieser würde unter bestimmten Bedingungen von den *prozeduralen Anforderungen* des Art. 10 KI-VO-E dispensieren, wenn die Qualität des *Ergebnisses* der KI-Modellierung den dafür aufgestellten Ansprüchen genügt.[60]

Technischer Hintergrund einer solchen Regelung wäre, dass gewisse Fehler in Trainingsdatenbeständen zu einem späteren Zeitpunkt der KI-Modellierung ausgeglichen werden können. So wird etwa im Bereich algorithmischer Fairness, also der Implementierung von Nichtdiskriminierungsmaßen in die Entwicklung von KI-Modellen, zwischen *pre-processing-*, *in-processing-* und *post-processing*-Verfahren

---

[57] Vgl. umfassend *Leyens*, Informationsintermediäre des Kapitalmarkts, 2017, 25 ff.

[58] Vgl. für die KI-Ethik *Hallensleben* et al., From Principles to Practice. An interdisciplinary framework to operationalise AI ethics, 2020, 26 ff.

[59] EG 3 und 5 KI-VO-E; *Europäische Kommission*, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz), COM/2021/206 final, 6.

[60] *Hacker*, 13 Law, Innovation and Technology 2021, 257 (300).

unterschieden.[61] Die Bereinigung der Trainingsdaten, etwa ihre Ausbalancierung, fällt unter die *pre-processing*-Variante.[62] Allerdings lassen sich dieselben Ergebnisse auch durch die anderen Verfahren erzielen, etwa indem die Ergebnisse der Modellierung, nach der jeweiligen Anwendung (*post-processing*), einem Korrekturalgorithmus unterzogen werden, der Unterschiede zwischen geschützten Gruppen ausgleicht.[63] Die Doppelung von Prozess- und Ergebnisregulierung hat zur Folge, dass Entwickler gewissermaßen auf *pre-processing*-Verfahren festgelegt werden, auch wenn *in-* oder *post-processing*-Verfahren möglicherweise ressourcenschonender oder performanzerhaltender eingesetzt werden könnten, um dasselbe Ergebnis zu erzielen. Diese Freiheit der effizienten Wahl der Fehlerbereinigung sollte durch den Einwand hinreichender Ergebnisqualität wiederhergestellt werden.

Konkret sollte der Einwand an zwei Voraussetzungen geknüpft werden. Erstens müsste die Einhaltung von Art. 10 KI-VO-E mit einem unverhältnismäßigen Aufwand einhergehen. Die Unverhältnismäßigkeit würde sich insbesondere auch daraus ergeben können, dass technisch im wesentlichen äquivalente Fehlerbereinigungsstrategien, die zu einem späteren Zeitpunkt der KI-Modellierung ansetzen, mit signifikant geringerem Aufwand durchführbar sind. Zweitens muss die hinreichende Performanz des KI-Systems hinsichtlich der Ergebnisse konkret nachgewiesen werden, vorzugsweise in repräsentativen Feldstudien (*pre-market testing*)[64] und nicht lediglich auf einem Testdatensatz, um die Beschränkungen der jeweiligen Testdatensätze zu überwinden.[65] Können die Entwickler die Voraussetzungen des Einwands darlegen und gegebenenfalls beweisen, so läge keine Verletzung von Art. 10 KI-VO-E vor. Damit ergibt sich letztlich ein aus der Zielsetzung der KI-VO-E abgeleiteter Primat der Ergebnis- vor der Prozessregulierung.

4. Partizipation: Co-Design von Trainingsdatensätzen

Ein zweites zentrales Ziel der europäischen KI-Strategie ist der Aufbau vertrauenswürdiger KI.[66] Zentral erscheint dabei nicht nur das Vertrauen der Nutzer und

---

[61] *Pessach/Shmueli*, Algorithmic Fairness, Working Paper 2020, arXiv preprint arXiv:2001.09784, 7 ff.; *Ntoutsi* et al., 10(3) Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2020), e1356, 3.

[62] *Pessach/Shmueli* (Fn. 61), 7.

[63] Siehe etwa *Feldman* et al., Certifying and removing disparate impact, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, 259; *Zehlike* et al., 34 Data Mining and Knowledge Discovery (2020), 163.

[64] Siehe dazu etwa *Geistfeld*, 105 Calif. L. Rev. (2017), 1611 (1678 ff.).

[65] Siehe *Northcutt* et al., Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks, Working Paper 2021, arXiv preprint arXiv:2103.14749; *Dehghani and others*, The Benchmark Lottery, Working Paper 2021, arXiv preprint arXiv:2107.07002.

[66] *Europäische Kommission*, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz), COM/2021/206 final, 6; High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019.

Marktteilnehmer in die Performanz der KI, sondern besonders auch die Einbeziehung von betroffenen Personengruppen.[67] Ein Kritikpunkt an dem KI-VO-E im Ganzen ist dabei, dass gerade die Perspektive derjenigen, auf die KI-Applikationen angewandt werden, fast vollständig ausgeblendet wird.[68] Dies zeigt sich nicht nur daran, dass der Verordnungsentwurf keine Betroffenenrechte vorsieht,[69] sondern auch daran, dass eine Intervention oder Diskussion der Betroffenen im Rahmen des Prozesses der Modellerstellung nicht vorgesehen ist. Vertrauen und Akzeptanz entstehen jedoch häufig gerade durch Partizipation.[70]

Insofern würden sich Regeln empfehlen, die Co-Designstrategien mit Blick auf die Erstellung und Kuratierung der Trainingsdaten vorsehen oder zumindest dazu anregen.[71] So könnten betroffene Personen etwa in die Entwicklung von diskriminierungssensitiven Datensätzen einbezogen werden. Beispiele dafür können die partizipatorische Evaluation der Repräsentativität und der Passgenauigkeit für anvisierte Gruppen nach Art. 10 Abs. 3 und Abs. 4 KI-VO-E sein. Durch entsprechende Workshops mit Stakeholdern und betroffenen Personen könnte ein inklusives und verantwortungsvolles ML-Design jenseits rein quantitativer Optimierung befördert werden. Zugleich kann eine partizipatorische Reflexion der Trainingsdaten Vorbild sein für Mitsprache und Kontestation von betroffenen Personen auch in anderen Bereichen der KI-Entwicklung.[72]

VI. Zusammenfassung

Trainingsdaten für KI-Systeme sind nicht nur in technischer Hinsicht zentral, sondern nehmen zunehmend auch einen Schwerpunkt in der regulatorischen Debatte ein. Eine besondere Herausforderung besteht hier in der Verschränkung informatorischer Konzepte und Maße mit rechtlichen Begriffen und Standards.

Das gegenwärtig geltende Recht offenbart hier zahlreiche Unzulänglichkeiten. Die Kriterien für Trainingsdatenqualität sind bestenfalls lückenhaft, die Anwen-

---

[67] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019, 14, 18.

[68] Eine Ausnahme findet sich etwa in den Informationspflichten nach Art. 52 KI-VO-E.

[69] *Veale/Zuiderveen Borgesius* (Fn. 51), Rn. 98; *Ebers* et al., 4(4) J (2021), 589 (600).

[70] Vgl. *Greenberg/Folger*, Procedural justice, participation, and the fair process effect in groups and organizations, in: Paulus (Hrsg.), Basic Group Processes, 1983, 235; *Tyler/Lind*, Procedural Justice, in: Sanders/Hamilton (Hrsg.), Handbook of Justice Research in Law, 65 (67).

[71] Siehe *Friedman/Hendry/Borning*, 11 Foundations and Trends in Human-Computer Interaction (2017), 63; *Passoth*, Die Demokratisierung des Digitalen, 424 Analysen & Argumente (2021), 1 (8 f.).

[72] Siehe etwa *Hildebrandt*, 20 Theoretical Inquiries in Law (2019), 83 (117 ff.); *Delacroix*, Journal of Cross-disciplinary Research in Computational Law 2022, 1; *Hacker/Passoth*, Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond, in: Holzinger et al. (Hrsg.), LNAI 13200: xxAI – Beyond Explainable AI, 2022, 343 (366 f.).

dung von DS-GVO und AGG einzelfallabhängig und häufig unklar. Demgegenüber sieht der Verordnungsentwurf für Künstliche Intelligenz in seinem Art. 10 eine umfassende Regelung der Daten-Governance vor, die allerdings nur bei Hochrisiko-Applikationen greift. Hier werden neben einigen klassischen Dimensionen von Datenqualität auch KI-spezifische Maße wie Repräsentativität und gruppenspezifische Ausgewogenheit eingeführt. Zudem ist zu begrüßen, dass ein eigenständiger datenschutzrechtlicher Erlaubnistatbestand für die Korrektur von potentiell diskriminierenden Verzerrungen in Trainingsdatensätzen geschaffen wird.

Allerdings bleiben auch etliche Kritikpunkte. So ist die Koordination von Art. 10 KI-VO-E mit den bestehenden Nichtdiskriminierungsregeln unzureichend. Friktionen können sich ferner aus der Doppelung einer Regulierung der KI-Ergebnisse über das allgemeine Haftungsrecht und Art. 15 KI-VO-E einerseits und einer Regulierung des Trainingsprozesses andererseits ergeben. Entwicklern wird damit möglicherweise verwehrt, effiziente Fehlerbereinigungsstrategien zu wählen. Schließlich gewähren die zum Teil vagen und nicht eindeutig einem spezifischen technischen Phänomen zuordenbaren Begriffe des Art. 10 KI-VO-E zwar Flexibilität, gehen aber mit erheblicher Rechtsunsicherheit einher, die angesichts der erheblichen Sanktionen von bis zu 6 % (nach KI-VO-E Juli 22: nur noch 4 %) des weltweiten Jahresumsatzes bei Verstoß gegen das Trainingsdatenregime dringend reduzierungsbedürftig erscheint.

Vor diesem Hintergrund unterbreitet der Beitrag vier Vorschläge für eine Weiterentwicklung des Trainingsdatenregimes. Erstens sollte man die Rechtssicherheit erhöhen. Es ließen sich beispielsweise *safe harbors* definieren, die je nach Risikoträchtigkeit der Anwendung bestimmte Fehlertoleranzen festlegen. Zweitens kann der Qualitätswettbewerb hinsichtlich Trainingsdaten womöglich von klaren Pflichtinformationen profitieren, wenn diese auf das jeweilige Zielpublikum zugeschnitten werden. Drittens sollte das Spannungsverhältnis zwischen Ergebnis- und Prozessregulierung so aufgelöst werden, dass eine Verletzung von Art. 10 KI-VO-E ausgeschlossen wird, sofern die Entwickler nachweisen können, dass sie etwaige Defizite in den Trainingsdaten an einem späteren Punkt der KI-Modellierung ausgeglichen und dadurch die Anforderungen an ein hinreichend optimiertes KI-Ergebnis erfüllt haben. Dies entspricht dem Ziel, effektive, performativ hochwertige KI zu fördern. Darüber sollte jedoch ein weiteres Ziel, jenes der vertrauenswürdigen KI, nicht in Vergessenheit geraten. Dem ließe sich etwa durch die partizipatorische Evaluation von Trainingsdaten in Co-Design-Workshops mit betroffenen Personen näherkommen. Solche Partizipationselemente könnten dann zugleich eine Blaupause für deren Integration in andere Bereiche der KI-Entwicklung darstellen.

KI-gestützte Kfz-Mobilität als Herausforderung für die Verbraucherpolitik

Eric Hilgendorf

I. Einleitung

Technik, so ist oft zu hören, muss dem Menschen dienen, und nicht umgekehrt. Damit wird ein wesentlicher Grundsatz sowohl der deutschen als auch der europäischen Verbraucherpolitik treffend auf den Punkt gebracht.[1] Um ihn durchzusetzen und womöglich auszubauen, müssen risikoträchtige technische Entwicklungen frühzeitig identifiziert und ethisch wie rechtlich begleitet werden: letzteres wird als (ethische und juristische) „Begleitforschung" bezeichnet.[2] Dazu gehört auch eine angemessene, rechtzeitig erfolgende Technikfolgenabschätzung.[3] Technik darf nicht nur an „internen" Parametern (Leistung, Effizenz i. e. S.) und ihrem Nutzen für Einzelne (Eigentümer, Aktienbesitzer) gemessen werden, auch die Folgen für das Gemeinwohl müssen in den Blick genommen werden. Dies gilt gerade auch für die neuen, auf Digitalisierung und KI beruhenden Mobilitätsformen.[4]

Alle Formen von Mobilität werden derzeit durch die drängenden Anforderungen des Klimaschutzes auf eine harte Probe gestellt. Weitere, damit zusammenhängende Herausforderungen bilden die Umstellung auf „eMobilität"[5] und die veränderte Wertschätzung des Autos, das vor allem in der jüngeren Generation häufig nicht mehr als wichtiges Statussymbol und Ausdruck individueller Freiheit betrachtet wird, sondern allenfalls noch als individuelles Fortbewegungsmittel Anerkennung genießt. Mobilität befindet sich deshalb heute in einer Phase radi-

---

[1] Das Konzept der „Verbraucherpolitik" wird hier in einem weiteren Sinne als das oft auf gesundheitliche und wirtschaftliche Belange beschränkte Konzept des „Verbraucherschutzes" verstanden. „Verbraucherpolitik" meint „alle Aktivitäten und Maßnahmen […] die auf die Verwirklichung von Verbraucherinteressen hinwirken", *Bala/Schuldzinski*, Verbraucherpolitik, in: Andersen u. a. (Hrsg.), Handwörterbuch des politischen Systems der Bundesrepublik Deutschland, 8. Aufl. 2021, S. 945 (946).

[2] Allgemein zu den Aufgaben des Technikrechts *Hilgendorf*, JZ 2012, 825 (827).

[3] *Grunwald*, Technikfolgenabschätzung – eine Einführung, 2. Aufl. 2010.

[4] Piallat (Hrsg.), Der Wert der Digitalisierung: Gemeinwohl in der digitalen Welt, 2021.

[5] Oekom e. V. – Verein für ökologische Kommunikation (Hrsg.), Postfossile Mobilität. Zukunftstauglich und vernetzt unterwegs, 2014; kritisch zur elektrischen Mobilität *Kneissl*, Die Mobilitätswende und ihre Brisanz für Gesellschaft und Weltwirtschaft, 2020, S. 124 ff.

kalen Umbruchs, einer disruptiven Veränderung, die wohl nur mit dem Aufkommen von PKWs zu Beginn des 20. Jahrhunderts verglichen werden kann.[6]

Gerade die deutsche Automobilindustrie wird in vielfacher Weise massiv herausgefordert. Sie bildet einerseits „das Rückgrat der deutschen Wirtschaft".[7] Wie keine andere Branche schafft sie in Deutschland Arbeitsplätze und leistet erhebliche Steuerzahlungen an den Fiskus, die zum Erhalt der sozialen Standards in der Bundesrepublik Deutschland dringend nötig sind.[8] Andererseits steht die Automobilindustrie in der Kritik, weil sie Umweltbelange vernachlässige,[9] die Wende zur eMobilität verschlafen und im sog. „Abgasskandal" sogar kriminelle Mittel angewendet habe, um ihre privilegierte Position zu verteidigen.[10] Auch wenn einige dieser Vorwürfe überzogen sein dürften, bleibt festzuhalten, dass die deutsche Automobilindustrie nicht bloß mit dem technologischem Umbruch zu kämpfen hat, sondern auch mit teilweise selbstverschuldeten Akzeptanzproblemen.

Eine weitere Herausforderung stellen neue Regulierungsvorschläge dar, welche derzeit auf EU-Ebene erarbeitet werden und die die Mobilität jedenfalls insofern betreffen, als sie KI-gestützt ist. Regelungen speziell für den Mobilitätsbereich werden ebenfalls bereits vorbereitet.[11] Sowohl die europäische Verkehrspolitik als auch die erst am Anfang stehende Regulierung von KI muss im Einklang mit den europäischen Verbraucherschutzvorgaben ausgestaltet werden, die nicht nur die Schaffung eines gemeinsamen Binnenmarktes befördern, sondern die europäi-

---

[6] Die verschiedenen Aspekte der gegenwärtigen „Mobilitätsdisruption" beleuchten neben den in Fn. 5 genannten Texten *Burns*, Autonomy. The Quest to Build the Driverless Car – and How It Will Reshape Our World, 2018; *Canzler/Knie*, Die digitale Mobilitätsrevolution. Vom Ende des Verkehrs, wie wir ihn kannten, 2016; *Dudenhofer*, Wer kriegt die Wende? Zeitenwende in der Autoindustrie, 2016; *Daum*, Das Auto im digitalen Kapitalismus. Wenn Algorithmen und Daten den Verkehr bestimmen, 2019; *Herrmann/Brenner*, Die autonome Revolution. Wie selbstfahrende Autos unsere Straßen erobern, 2018; *Kahle*, Mobilität in Bewegung. Wie soziale Innovationen unsere mobile Zukunft revolutionieren, 2021; *Lipson/Kurman*, Driverless. Intelligent Cars and the Road Ahead, 2016; Maurer u. a. (Hrsg.), Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte, 2015; *Rammler*, Schubumkehr. Die Zukunft der Mobilität, 2014; W.I.R.E (Hrsg.), Transforming Transport. Zur Vision einer intelligenten Mobilität, 2016; aufschlußreich auch acatech (Hrsg.), Horizonte: Transformation der Mobilität, 2021.

[7] *Kneissl*, Die Mobilitätswende (Fn. 5), S. 137.

[8] Es ist eine leider immer wieder übersehene, wenngleich selten offen geleugnete Binsenweisheit, dass ein leistungsfähiger Sozialstaat eine leistungsfähige Wirtschaft voraussetzt, in der die notwendigen Mittel erarbeitet werden. Umgekehrt gilt aber auch, dass die Wirtschaft von einem hohen allgemeinen Bildungsgrad, sozialem Ausgleich, engagierter Verbraucherpolitik und sozialer Mobilität erheblich profitiert.

[9] Aufschlußreich oekom e.V. – Verein für ökologische Kommunikation (Hrsg.), Mobilitätswende. Die Zeit ist reif, 2021.

[10] Zum Abgasskandal: *Borgeest*, Manipulation von Abgaswerten: Technische, gesundheitliche, rechtliche und politische Hintergründe des Abgasskandals, 2. Aufl. 2021; *Ring*, Straßenverkehrsrecht Bd. 20 (2020), S. 401 ff.; *ders.*, Straßenverkehrsrecht Bd. 21 (2021), S. 121 ff.; siehe auch *Kneissl*, Die Mobilitätswende (Fn. 5), S. 121 ff.

[11] Für einen Überblick zu den EU-Aktivitäten im Bereich des automatisierten und vernetzten Fahrens siehe https://www.europarl.europa.eu/news/en/headlines/economy/20190110STO23102/self-driving-cars-in-the-eu-from-science-fiction-to-reality.

schen Verbraucher vor unnötigen Risiken bewahren und angemessene Verbraucherrechte sicherstellen sollen.[12] Nicht nur auf der nationalen, sondern auch auf der europäischen Ebene ist Verbraucherpolitik mithin eine Querschnittsaufgabe, die sich auch und gerade bei der Regulierung von moderner Mobilität stellt.[13] Es steht zu erwarten, dass der geltende nationale Rechtsrahmen für das automatisierte und vernetzte Fahren[14] in Zukunft immer stärker von europäischen Vorgaben überlagert und beeinflusst werden wird.

Dazu gehören nicht bloß automatisierte und vernetzte Straßenfahrzeuge, obgleich das damit angesprochene „autonome Fahrzeug" häufig im Mittelpunkt des Medieninteresses steht. Zur modernen Mobilität gehören auch der Schienen- und der Luftverkehr, der Verkehr auf dem Wasser und schließlich auch die Nutzung der Straße durch Motorradfahrer, Radfahrer und Fußgänger.[15] Die Mobilität der Zukunft wird wohl vor allem von vernetzten Verkehrssystemen[16] geprägt sein, gerade in den Ballungsräumen: Nach der Anreise mit dem Auto, dem Bus oder der Bahn steigt der Fahrgast in den innerstädtischen Schienenverkehr um, der ihn bis in die unmittelbare Nähe seines Zieles bringt. Der „letzte" Kilometer mag mit einem autonom fahrenden Taxi, einem bereitstehenden Pedelec oder Fahrrad oder auch zu Fuß zurückgelegt werden. Vernetzte Verkehrssysteme bedürfen genauester Abstimmung und Kontrolle; sie sind heute nur noch digitalisiert und KI-gesteuert vorstellbar. Die Bereitstellung derartiger Systeme, die nicht bloß bequem und reibungslos[17] funktionieren, sondern auch den Vorgaben eines nachhaltigen, energiesparenden und umweltschützenden Verkehrs genügen sollen,

---

[12] Aussagekräftiger Überblick zu den europäischen Vorgaben der Verbraucherpolitik bei *Pieper*, in: Bergmann (Hrsg.), Handlexikon der Europäischen Union, 6. Aufl. 2022, S. 1045 f., kritisch *Luger*, ZEUP 2018, 788 ff.

[13] Verkehrspolitik war freilich immer schon mit Belangen des Gemeinwohls und damit auch dem Umwelt- und dem Verbraucherschutz verknüpft, selbst wenn Konzepte wie „Verbraucherpolitik" oder „Nachhaltigkeit" jüngeren Datums sind, dazu etwa *Ammoser*, Das Buch vom Verkehr. Die faszinierende Welt von Mobilität und Logistik, 2014, S. 314 ff. *Merki*, Verkehrsgeschichte und Mobilität, 2008, S. 88 ff. unterscheidet folgende Nachhaltigkeitsdimensionen des modernen Verkehrs: soziale Kosten, Energieverbrauch, Landverschleiß, und privat vs. staatlich. Zu Geschichte der Mobilität allgemein *Schreiber*, Verkehr, 1969.

[14] Problemaufriss bei *Hilgendorf*, Automatisiertes Fahren und Recht. Gutachten für den 53. Verkehrsgerichtstag in Goslar, in: 53. Deutscher Verkehrsgerichtstag 2015, S. 55 ff., ausführlich Oppermann/Stender-Vorwachs (Hrsg.), Autonomes Fahren. Technische Grundlagen, Rechtsprobleme, Rechtsfolgen, 2. Aufl. 2019; zu strafrechtlichen Problemen auch *Nestler*, Jura 2021, S. 1183 ff., zu Grundrechtsgefährdungen Roßnagel/Hornung (Hrsg.), Grundrechtsschutz im Smart Car. Kommunikation, Sicherheit und Datenschutz im vernetzten Fahrzeug, 2019; siehe schließlich auch das Schwerpunktheft „Verkehrswende" der Zeitschrift für das Recht der Digitalen Wirtschaft (ZdiW) 2022, Heft 1.

[15] Verkehrspolitik muss deshalb viel mehr in den Blick nehmen als nur den autonomen Straßenverkehr, dazu etwa *Ammoser*, Das Buch vom Verkehr (Fn. 6), S. 241 ff.

[16] Acatech, Transformation der Mobilität (Fn. 5), S. 22 ff., Lemmer (Hrsg.), Neue autoMobilität II: Kooperativer Straßenverkehr und intelligente Verkehrssteuerung für die Mobilität der Zukunft, 2019 (acatech Studie).

[17] Dazu gehört immer mehr auch die Widerständigkeit (oder: Resilienz) gegenüber Cyberkriminalität.

ist die vielleicht größte Herausforderung der Verkehrspolitik der nächsten Jahrzehnte.[18] In diesem Zusammenhang stellen sich neben praktischen Herausforderungen auch schwierige, gerade auch aus Verbrauchersicht wichtige Grundlagenprobleme, etwa der Ausgleich zwischen individueller Freiheit und (rechtlich oder technisch) erzwungener Regelkonformität.[19]

Jedenfalls im Hinblick auf die Regulierung des automatisierten Fahrens ist Deutschland in Vorleistung gegangen und hat im Sommer 2021 als erstes Land der Welt Regeln für den autonomen Betrieb von Kraftfahrzeugen erlassen. Dem gingen langjährige Planungen, ethische und rechtliche Analysen sowie praktische Tests voraus. Die Reform ist nach bisherigem Erfahrungs- und Diskussionsstand gelungen,[20] so dass zu erwarten ist, dass das deutsche Regelungsmodell auch auf die zu erwartende Europäische Regelung des automatisierten und autonomen Fahrens Einfluss haben wird. Derzeit (Februar 2022) werden erste praktische Projekte auf der Grundlage der neuen Gesetzeslage vorbereitet.[21]

II. Zentrale Elemente der Verbraucherpolitik im Kontext moderner Mobilität

Eine gemeinwohlorientierte (und tatsächlich gemeinwohlfördernde) Verbraucherpolitik im Feld der modernen Mobilität lässt sich u. a. anhand folgender Leitplanken entwickeln:

- Mobilität sollte nicht allein als Ausdruck individuellen Freiheitsstrebens gesehen werden, sondern als wesentlicher Bestandteil des Gemeinwohls.[22]
- Alle individuellen Mobilitätsziele sollten möglichst rasch und bequem erreicht werden können. Dafür muss ein Kompromiss zwischen verschiedenen frei wählbaren Mobilitätsformen (z. B. Auto oder Fahrrad) gefunden werden.[23]
- Mobilität in allen ihren Formen muss so sicher wie möglich gestaltet sein. Dabei muss zwischen individueller Freiheit und risikominimierender Gleichheit abgewogen werden.[24] Gerade unter Verbraucherschutzgesichtspunkten ist Risikominderung durch technischen Zwang (technologischer Paternalismus) nicht per se negativ zu bewerten.[25]

---

[18] *Rammler*, Schubumkehr (Fn. 6), S. 73 ff.

[19] Dazu *Birnbacher*, Fahrerlose Fahrzeuge – wieviel Gleichheit, wieviel Freiheit?, in: Beck/Kusche/Valerius (Hrsg.), Digitalisierung, Automatisierung, KI und Recht, 2020, S. 17 ff.

[20] *Hilgendorf*, JZ 2021, 444 (454).

[21] Siehe *Buchter/Tatje*, in: DIE ZEIT vom 10.2.2022, S. 24.

[22] *Kahle*, Mobilität in Bewegung (Fn. 6), S. 23.

[23] Dies bedeutet faktisch, dass die Privilegierung des Automobilverkehrs zulasten der Fussgänger und Radfahrer zurückgedrängt werden muss. Dieser Prozess ist bereits im Gange. Einen Blick in die (mögliche) Zukunft gewährt *Rammler*, Schubumkehr (Fn. 6), S. 219 ff.

[24] Siehe den Nachweis oben Fn. 19.

[25] *Hilgendorf*, Gemeinwohlorientierte Gesetzgebung auf Basis der Vorschläge der EU „High-Level-Expert-Group on Artificial Intelligence", in: Piallat, Wert der Digitalisierung (Fn. 4), S. 223 (247 ff.).

- Die Kosten der Mobilität sollten für alle tragbar sein, d. h. bequeme Mobilität sollte nicht zu einem Privileg der Reichen werden. In diesen Zusammenhang gehört auch die Option eines kostenfrei nutzbaren ÖPNV.[26]
- Alle Formen der Mobilität sollten möglichst umweltschonend und nachhaltig ausgestaltet sein.[27] Dafür sind technische Innovationen nötig.[28]
- Soweit Mobilität die Bedienung von Geräten voraussetzt (z. B. das Bestellen eines autonomen Shuttles, das Steuern eines Fahrzeugs), sollte die Technik so gestaltet sein, dass sie der Durchschnittsverbraucher ohne Weiteres nutzen kann.[29]
- Auch für ältere und behinderte Menschen müssen angemessene Mobilitätsangebote zur Verfügung stehen.[30]
- Der Wechsel zwischen unterschiedlichen Mobilitätsformen (z. B. das Umsteigen von Bahn zu Bus oder Shuttle) muss grundsätzlich für alle problemlos möglich sein.
- Im öffentlichen Personenverkehr müssen angemessene Möglichkeiten angeboten werden, um im Verkehr aufgetretene Probleme zu melden, Verbesserungsvorschläge vorzubringen und evtl. Kompensationen zu beantragen. Es sollten hinreichend Möglichkeiten zur Verfügung stehen, um bürgerschaftliches Engagement zu unterstützen und in der Verkehrsplanung einzubinden.
- Das Verkehrsrecht sollte so gestaltet sein, dass Betroffenen im Schadensfall unkompliziert und unbürokratisch Ersatz geleistet werden kann. Der zunehmende Einsatz von KI und anderen Formen vernetzter Technik darf nicht zu Haftungslücken oder Verantwortungsdiffusion führen.[31]
- Privaten Verbraucherschutzorganisationen, die sich im Bereich Mobilität engagieren, müssen angemessene Beteiligungsmöglichkeiten bei der Planung und Umsetzung von Projekten gewährt werden. Auch im Rahmen von Gesetzgebungsverfahren sind sie zu beteiligen.[32]

---

[26] Zur Verbilligung von Verkehrsdienstleistungen allgemein *Merki*, Verkehrsgeschichte und Mobilität (Fn. 13), S. 81 ff.

[27] *Rammler*, Schubumkehr (Fn. 6), S. 75 ff.

[28] Ein ganzes Panorama zukunftsweisender Ideen entfaltet *Rammler*, Schubumkehr (Fn. 6), S. 219 ff.

[29] Das Problemfeld des „digital divide" wird in der Debatte um die neue Mobilität regelmäßig vernachlässigt. Vor allem die Belange älterer Menschen, die sich durch die Geschwindigkeit der „digitalen Revolution" überfordert fühlen, werden regelmäßig vernachlässigt. Nach wie vor gibt es in Deutschland zahlreiche Menschen, die nicht über ein Smartphone verfügen, und es existiert auch (noch?) keine Rechtspflicht, ein Smartphone anzuschaffen und bei sich zu führen.

[30] Zu den damit verbundenen Problemstellungen und Herausforderungen *Stöppler*, Menschen mit (Mobilitäts-)Behinderung. Teilhabe und Verkehrssicherheit. Handbuch für Fachkräfte zur Förderung der Mobilitätskompetenzen von Menschen mit Behinderungen, 2015 (Deutscher Verkehrssicherheitsrat, Schriftenreihe Verkehrssicherheit 18).

[31] Allgemein zum Thema „Verantwortung im Straßenverkehr" *Hilgendorf*, in Roßnagel/Hornung, Grundrechtsschutz im Smart Car (Fn. 14), S. 147 ff.

[32] Bekannte und einflussreiche Organisationen dieser Art sind etwa der Allgemeine Deutsche Automobil-Club (ADAC) oder der Deutsche Verkehrssicherheitsrat (DVR).

III. Ethische und rechtliche Vorgaben für Künstliche Intelligenz auf EU-Ebene – ein Überblick

Vorschläge für die ethische und rechtliche Rahmung und Einhegung Künstlicher Intelligenz existieren zuhauf.[33] Von besonderer Bedeutung sind die Regulierungsansätze, die auf Europäischer Ebene entwickelt wurden. Im Jahr 2018 wurde eine „Hochrangige Expertengruppe", die „European High Level Expert Group on AI" (HLEG AI) eingesetzt, welche in mühevoller zweijähriger Arbeit Grundlagen für eine Regulierung von KI in Europa und darüber hinaus erarbeitet hat.[34] Zu den Ergebnissen zählen insbesondere zwei Arbeitspapiere, die „Ethics Guidelines for trustworthy AI" (8.4.2019)[35]und die „Trustworthy AI Assessment List" (ALTAI, 17.7.2020).[36] Das erste Papier widmet sich den Grundlagen, während im zweiten Anwendungsszenarien dargestellt und Risikosituationen strukturiert und analysiert werden.

Ausgehend von der Menschenwürdegarantie (Art. 1 GRCH, Art. 1 GG) entfaltet der erste Text die Idee einer „vertrauenswürdigen" („trustworthy") KI in mehreren Stufen.[37] Dabei wird die Bedeutung einer Orientierung am Menschen und seinen Bedürfnissen („human-centric approach") betont. Die Regelungsansätze der HLEG sind deshalb mit den Grundlagen des Verbraucherschutzes, wie sie gerade in Deutschland seit den 60er Jahren erarbeitet und umgesetzt wurden, in Einklang zu bringen;[38] sie gehen teilweise sogar noch deutlich darüber hinaus.

Im 2020 publizierten Weißbuch der EU „On Artificial Intelligence – A European Approach to Excellence and Trust"[39] wurden die Vorschläge der HLEG AI aufgegriffen und weiter ausdifferenziert. So entstand die Grundlage für den Entwurf einer EU-Richtlinie zur Regulierung Künstlicher Intelligenz, der im April 2021 publiziert wurde.[40] Flankiert wird dieser Vorstoß von Vorschlägen zu neuen

---

[33] Guter Überblick bei *Jobin/Ienka/Vayena*, Nature Machine Intelligence 2019, 389 ff.

[34] Dazu näher *Hilgendorf*, Gemeinwohlorientierte Gesetzgebung auf Basis der Vorschläge der EU „High-Level-Expert-Group on Artificial Intelligence", in Piallat, Wert der Digitalisierung (Fn. 4), S. 223 (228 f.).

[35] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[36] file:///C:/Users/strp039/AppData/Local/Temp/altai_final_14072020_cs_accessible2_jsd5pdf_correct-title_3AC24743-DE11-0B7C-7C891D1484944E0A_68342.pdf.

[37] *Hilgendorf*, Gemeinwohlorientierte Gesetzgebung auf Basis der Vorschläge der EU „High-Level-Expert-Group on Artificial Intelligence", in Piallat, Wert der Digitalisierung (Fn. 4), S. 223 (229 ff.).

[38] Der Zusammenhang wird besonders deutlich, wenn man Verbraucherpolitik als Grundrechtsaktivierung versteht und dadurch letztlich an die Menschenwürde rückbindet.

[39] COM (2020) 65 final vom 19.2.2020.

[40] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislatice acts, COM/2021/206 final.

europäischen Regelungsvorhaben im Bereich der Datenwirtschaft und zum Datenschutz, vor allem dem „Data Governance Act“[41] und dem „Data Act“.[42]

IV. Auf dem Weg zu einem Europäischen „Gesetz über Künstliche Intelligenz“[43]

Der „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“[44] ist nicht weniger als der Versuch einer weltweit ersten umfassenden rechtliche Regulierung Künstlicher Intelligenz. Der Vorschlag ist auf breites Interesse gestoßen, wobei sich befürwortende und kritische Stimmen in etwa die Waage halten.[45] Im Folgenden sollen die Kernelemente des Vorschlags in allgemeiner Perspektive herausgearbeitet und einige zentrale pro und contra-Gesichtspunkte dargelegt werden.

Die Kommission begründet ihren Vorschlag damit, dass der „Einsatz Künstlicher Intelligenz zur Verbesserung von Prognosen, zur Optimierung von Abläufen und der Ressourcenzuweisung sowie zur Personalisierung der Diensteerbringung … für die Gesellschaft und die Umwelt von Nutzen sein und Unternehmen sowie der europäischen Wirtschaft Wettbewerbsvorteile verschaffen“ könne. Insbesondere in den Sektoren „Klimaschutz, Umwelt und Gesundheit, öffentlicher Sektor, Finanzen, Mobilität (!, E. H.), Inneres und Landwirtschaft“ könne KI eingesetzt werden. Dem zu erwartenden sozioökonomischen Nutzen stünden allerdings „neue Risiken oder Nachteile für den Einzelnen oder die Gesellschaft“ gegenüber. Deswegen sei eine Regulierung von KI und ihrer Anwendungen erforderlich.[46]

1. Ziel der Verordnung; verbotene Technologien

Ziel der geplanten Verordnung ist es, „harmonisierte Vorschriften für das Inverkehrbringen, die Inbetriebnahme und die Verwendung von Systemen der künstlichen Intelligenz“ in der Europäischen Union zu schaffen (Art. 1a). Bestimmte

---

[41] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM/2020/767 final.

[42] Der Data Act wird derzeit noch vorbereitet. Eine öffentliche Konsultation hat bereits stattgefunden, https://data.europa.eu/en/news/public-consultation-data-act. Ende Februar 2022 wurde ein erster Entwurf des „Data Acts“ publiziert.

[43] Die nachfolgenden Passagen sind einem von mir verfassten Papier des Bayerischen KI-Rats zum neuen EU-Regulierungsvorschlag entnommen.

[44] https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

[45] In Auswahl: *Bomhard*, Recht digital 2021, 276 ff.; *Ebert/Spieker gen. Döhmann*, NVwZ 2021, 1188 ff:, *Engelmann*, Recht digital 2021, 317 ff.; *Roos*, MMR 2021, 844 ff.

[46] VO-Entwurf, S. 1.

Formen von KI sollen ganz verboten, für Hochrisiko-KI-Systeme besondere Anforderungen formuliert werden (Art. 1b, c). Der Anwendungsbereich der Verordnung soll sehr weit sein: sie soll für alle Anbieter gelten, „die KI-Systeme in der Union in Verkehr bringen oder in Betrieb nehmen, unabhängig davon, ob diese Anbieter in der Union oder in einem Drittland niedergelassen sind". Der Anwendungsbereich wird auch dadurch eröffnet, dass sich die Nutzer der KI Systeme in der Union befinden (Art. 2).

Art. 3 Nr. 1 definiert ein KI-System als „eine Software, die mit einer oder mehreren der in Anhang I aufgeführten Techniken und Konzepten entwickelt worden ist und im Hinblick auf eine Reihe von Zielen, die vom Menschen festgelegt werden, Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder Entscheidungen vorbringen kann, die das Umfeld beeinflussen, mit dem sie interagieren". Als entsprechende Techniken werden in Anhang I Konzepte des maschinellen Lernens genannt, des Weiteren Logik und wissenschaftsgestützte Konzepte sowie statistische Ansätze. Das dem Regelungsvorschlag zugrundeliegende KI-Verständnis ist damit außerordentlich weit, möglicherweise zu weit, denn es dürfte heute kaum noch anspruchsvolle Software geben, die nicht auf den erwähnten Techniken des maschinellen Lernens, der Logik, auf „wissenschaftsgestützten Konzepten" oder „statistischen Ansätzen" beruht.[47]

Gänzlich verboten sollen folgende Praktiken werden: „das Inverkehrbringen, die Inbetriebnahme oder die Verwendung von KI Systemen", die „Techniken der unterschwelligen Beeinflussung außerhalb des Bewusstseins einer Person" einsetzen, wenn dadurch einer Person physischer oder psychischer Schaden zugefügt werden kann (Art. 5 Abs. 1a). Verboten sind des Weiteren Systeme, die eine besondere Schwäche oder Schutzbedürftigkeit einer bestimmten Gruppe von Personen ausnutzen, „um das Verhalten einer dieser Gruppe angehörenden Person in einer Weise wesentlich zu beeinflussen", die einer Person schaden kann (Art. 5 Abs. 1b). Auch die staatliche Einrichtung von „social-scoring"-Systemen nach Chinesischem Muster[48] sollen verboten werden (Art. 5 Abs. 1c). Eine vierte Gruppe verbotener Praktiken sind biometrische Echtzeit-Fernidentifizierungssysteme, die in öffentlich zugänglichen Räumen zur Strafverfolgung eingesetzt werden (Art. 5 Abs. 1d), worunter insbesondere Systeme zur Gesichtserkennung zu verstehen sein dürften.[49]

---

[47] Zu anderen KI-Konzepten siehe *Boden*, AI-Its nature and future, 2016, 1 ff.

[48] Guter Überblick bei *Fischer*, Das chinesische Sozialkreditsystem. Durch Vertrauen zu Innovation und Wettbewerbsfähigkeit? Universitas 76 (2021), S. 71 ff.

[49] Ausnahmen bilden die gezielte Suche nach „bestimmten potentiellen Opfern von Straftaten" oder nach „vermissten Kindern", das „Abwenden einer konkreten, erheblichen und unmittelbaren Gefahr für das Leben oder die körperliche Unversehrtheit natürlicher Personen oder eines Terroranschlags", sowie das „Erkennen, Aufspüren, Identifizieren oder Verfolgen eines Täters oder Verdächtigen" einer besonders schweren Straftat.

2. Hochrisiko-Systeme und ihre Rahmenbedingungen

Den mit Abstand größten Raum nimmt die Regelung der Hochrisiko-KI-Systeme ein.[50] Dazu sollen nach Anhang III folgende Systeme gehören: Systeme, die für die biometrische Echtzeit- Fernidentifizierung natürlicher Personen verwendet werden (Nr. 1), Systeme, die bestimmungsgemäß als Sicherheitskomponenten in der Verwaltung und dem Betrieb des Straßenverkehrs sowie in der Wasser-, Gas-, Wärme- und Stromversorgung eingesetzt werden sollen (Nr. 2), Systeme, die für Entscheidungen über den Zugang oder die Zuweisung natürlicher Personen zu Einrichtungen der allgemeinen und beruflichen Bildung verwendet werden sollen (Nr. 3) und Systeme im Zusammenhang mit der Beschäftigung, dem Personalmanagement und dem Zugang zur Selbstständigkeit, insbesondere Systeme im Zusammenhang mit Bewerbungen und Einstellungsverfahren (Nr. 4). Erfasst sind ferner Systeme, die für grundlegende private und öffentliche Dienste und Leistungen eingesetzt werden, etwa im Zusammenhang mit öffentlichen Unterstützungsleistungen, Verfahren der Kreditwürdigkeitsprüfung oder der Entsendung von Not- und Rettungsdiensten (Nr. 5).

Zu den Hochrisiko-Systemen gehören dem Entwurf zufolge auch Systeme, die im Zusammenhang mit Strafverfolgung, auch und gerade im Ermittlungsverfahren, eingesetzt werden (Nr. 6), des Weiteren Systeme, die im Zusammenhang mit Migration, Asyl und Grenzkontrolle Verwendung finden sollen (Nr. 7). Zu den Hochrisiko-Systemen zählen auch KI-Systeme, „die bestimmungsgemäß Justizbehörden bei der Ermittlung und Auslegung von Sachverhalten und Rechtsvorschriften und bei der Anwendung des Rechts auf konkrete Sachverhalte unterstützen sollen" (Nr. 8). Schließlich soll es sich um ein Hochrisiko-KI-System handeln, wenn das System als Sicherungskomponente eines Produkts verwendet werden soll, das unter bestimmte, in Anhang II des VO-E. näher aufgeführte Harmonisierungsrechtsvorschriften der Union fällt (Art. 6 Abs. 1). In Art. 7 wird der EU-Kommission die Befugnis übertragen, delegierte Rechtsakte zur Änderung der Liste in Anhang III zu erlassen, das heißt neue Hochrisiko- KI-Systeme zu definieren.

Die anderen KI-Systeme gelten als Systeme mit geringem oder minimalem Risiko, denen keine besonderen Anforderungen auferlegt werden. Stattdessen wird auf Transparenz (Art. 52) und Selbstregulierung via Verhaltenskodizes (Art. 69) verwiesen.

Für die Hochrisiko-KI-Systeme soll eine ganze Reihe von Anforderungen gelten, die in Kapitel II des Entwurfs näher festgelegt werden. Dazu gehört etwa die Einrichtung eines Risikomanagementsystems (Art. 9), besondere Verfahren der Daten-Kontrolle und Daten- Governance, das Erfordernis einer technischen Dokumentation (Art. 11) bestimmte Aufzeichnungspflichten (Art. 12), die Sicherstellung von Transparenz und die Bereitstellung von Informationen für die Nutzer

[50] Titel III Kap. 1–5.

(Art. 13) und die Sicherstellung menschlicher Aufsicht (Art. 14). Hinzu tritt das Gebot von Genauigkeit, Robustheit und Cyber-Sicherheit (Art. 50).

Die skizzierten Pflichten obliegen zunächst den Anbietern der Hochrisiko-Systeme (Art. 16), die außerdem ein Qualitätsmanagementsystem einrichten müssen (Art. 17). Vorgeschrieben sind auch eine technische Dokumentation (Art. 18), die Durchführung von Konformitätsbewertungsverfahren (Art. 19) und automatisch erzeugte Protokolle (Art. 20). Grundsätzlich dieselben Pflichten treffen den Produkthersteller (Art. 24). Auch die Importeure von KI-Systemen (Art. 26) und Händler (Art. 27) werden in die Pflicht genommen. Dasselbe gilt für die Nutzer (Art. 29), die insbesondere verpflichtet werden, KI-Systeme nur entsprechend der dem System beigefügten Gebrauchsanweisungen zu nutzen (Art. 29 Abs. 1). In jedem Mitgliedstaat müssen notifizierende Behörden geschaffen werden, also Behörden, die für die Einrichtung und Durchführung der erforderlichen Verfahren zu Bewertung, Benennung und Modifizierung von Konformitätsbewertungsstellen und deren Überwachung zuständig sind. Näheres ist in dem außerordentlich umfangreichen und komplexen vierten Kapitel des Entwurfs geregelt.

In Art. 43 wird die aufwändige Konformitätsbewertung behandelt, die für KI- Systeme vorgenommen werden muss. Auch die damit zusammenhängenden Fragen werden in großem Detail reguliert. In Art. 52 werden für bestimmte KI-Systeme besondere Transparenzpflichten festgelegt. So soll ein Mensch erkennen können, dass er es mit einem KI-System zu tun hat, sofern dies nach den Umständen nicht offensichtlich ist (Art. 52 Abs. 1)

In den Art. 53 ff. werden Maßnahmen zur Innovationsförderung behandelt.[51] Dazu gehören etwa KI-Real-Labore (Art. 53). Nach Art. 56 soll ein Europäischer Ausschuss für Künstliche Intelligenz geschaffen werden, der die Kommission mit Blick auf die KI-Regulierung berät und unterstützt. Des Weiteren sollen in jedem Mitgliedstaat nationale Aufsichtsbehörden benannt werden. Auch nach dem Inverkehrbringen obliegen insbesondere den Anbietern noch bestimmte Beobachtungs- und Überwachungspflichten. Um die Einhaltung der VO durchzusetzen, sieht Art. 71 erhebliche Sanktionen für den Fall einer Zuwiderhandlung vor. In bestimmten Fällen können Verstöße mit Geldbußen bis zu 30 Millionen Euro oder (bei Unternehmen) von bis zu 6 % des gesamten weltweiten Jahresumsatzes des vergangenen Geschäftsjahres verhängt werden (Art. 71 Abs. 3).

3. Vorläufige Bewertung

Was bedeutet dies alles für die KI-gestützte Mobilität? Der Versuch, KI-Systeme umfassend zu regulieren, erscheint grundsätzlich begrüßenswert. Um derartige Systeme am Markt zu platzieren, sind Rechtssicherheit und Verbrauchervertrauen

[51] Dahinter steht die Erkenntnis, dass Technikregulierung durchaus auch innovationsfördernde Wirkungen haben kann, wenn sie gut durchdacht und sinnvoll umgesetzt wird,

und damit ein klarer und nachvollziehbar rechtlicher Rahmen erforderlich. Es scheint jedoch, dass der Entwurf an nicht wenigen Stellen über die selbst gesetzten Ziele hinausschießt.

Das dem Entwurf zugrunde gelegte Verständnis von KI-Systemen ist außerordentlich weit, so das zahlreiche Unternehmen, nicht nur, aber gerade auch im Bereich der Mobilität betroffen sein dürften.[52] Bedenken erregt dies insbesondere deshalb, weil die Regulierung gerade der Hochrisiko-KI-Systeme außerordentlich komplex, bürokratieanfällig und für die Unternehmen belastend ist. Dass die Europäische Kommission ermächtigt werden soll, die Liste von Hochrisiko-Systemen einseitig auszuweiten (Art. 7), ist mit dem selbsterklärten Ziel von Rechtssicherheit nicht zu vereinbaren. Problematisch ist auch das Verhältnis zur Forschungsfreiheit, Art. 13 Satz 1 GRCH, Art. 5 Abs. III GG.

Nicht überzeugend sind des Weiteren die zahlreichen Vorgaben für Hersteller und Anbieter von Hochrisiko-Systemen, die sich über weite Strecken mit den Vorgaben von Produkthaftung überschneiden und zu unklaren Doppelungen führen. Für den Bereich der PKW-Mobilität, deren KI-Systeme als Hochrisiko-Systeme einzustufen sein dürften (Art. 6 Abs. 1 des Entwurfs i. V. m. Anhang II Abschnitt B Nr. 6) existieren insbesondere bei der Typzulassung bereits Zulassungsvorgaben auf EU-Ebene, deren Verhältnis zu den neuen, für Hoch-Risiko-Systeme vorgeschriebenen Vorgaben zu klären wäre. Auch das Verhältnis der geplanten Verordnung zur DSGVO wirft Fragen auf.[53]

Einige der Vorgaben der Regulierung sind schlichtweg nicht erfüllbar. So heißt es etwa in Art. 10 Abs. 3, dass die für das Training von KI-Systemen verwendeten Daten frei von Irrtümern sein müssen. Dies ist gar nicht möglich; vielmehr ist davon auszugehen, dass Datensätze stets auch Fehler enthalten. Bedenken erregt des Weiteren Art. 29, der auch Nutzern von Hochrisiko-KI-Systemen erhebliche Pflichten auferlegt. Diese Regelung erscheint wenig verbraucherfreundlich. Stattdessen würde es ausreichen, auf das bisherige, schon sehr leistungsstarke Produkthaftungsrecht zu verweisen. Alles in allem wird man den „ersten Aufschlag" zur Regulierung von KI in Europa als nur teilweise gelungen einstufen können.

Der Verordnungsvorschlag enthält keine Regeln speziell für den Mobilitätsbereich; vielmehr bildet der KI-gestützte Verkehr nur einen (allerdings besonders wichtigen) Anwendungsbereich des Regelungsentwurfs. Immerhin werden in Anhang III (Hochrisiko-KI-Systeme gemäß Artikel 6 Abs. 2) unter 2a) KI-Systeme im Straßenverkehr noch einmal ausdrücklich aufgelistet. Dieser streng systematisch auf allgemeine Regeln konzentrierte Regelungsansatz dürfte dem Problembereich angemessen sein, denn die Art und Zahl der konkreten Anwendungsfälle von KI lässt sich heute kaum seriös abschätzen.

---

[52] Siehe oben IV.2. und 3.

[53] Es handelt sich hierbei um ein Problem, das sich auch beim neuen Data Act und vielen anderen neuen Digitalgesetzen der EU zeigt.

V. Die Reform des deutschen Straßenverkehrsrechts im Sommer 2021 – ein erster nationaler Rechtsrahmen für KI-gestützte Mobilität

1. Grundlagen

Im Sommer 2021 wurde das deutsche Straßenverkehrsgesetz um Regelungen für Fahrzeuge der Stufe vier[54] ergänzt.[55] Grundlage der neuen Regulierung ist § 1d StVG, der den Betrieb von Kraftfahrzeugen mit autonomer Farbfunktion in festgelegten Betriebsbereichen regelt. Darin werden vier entscheidende Konzepte definiert:

Nach § 1d Abs. 1 handelt es sich um ein Kraftfahrzeug mit autonomer Fahrfunktion, wenn (1) das Kraftfahrzeug „die Fahraufgabe ohne eine fahrzeugführende Person selbstständig in einem festgelegten Betriebsbereich erfüllen kann" und das Fahrzeug außerdem (2) über eine technische Ausrüstung verfügt, deren Art in § 1e Abs. 2 des neuen Gesetzes näher festgelegt ist. Ein „festgelegter Betriebsbereich" ist der „örtlich und räumlich bestimmte öffentliche Straßenraum, in dem ein Fahrzeug mit autonomer Fahrfunktion betrieben werden darf." Ein drittes Konzept der neuen Regulierung ist das Konzept der „technischen Aufsicht". Es soll sich hierbei um eine natürliche Person handeln, die das Kraftfahrzeug „während des Betriebs deaktivieren und für das Fahrzeug Fahrmanöver freigeben kann".

Abs. 4 legt das Konzept des „risikominimalen Zustands" fest. „Risikominimaler Zustand im Sinne dieses Gesetzes ist ein Zustand, in dem sich das Kraftfahrzeug mit autonomer Fahrfunktion auf eigene Veranlassung oder auf Veranlassung der technischen Aufsicht an einer möglichst sicheren Stelle in einem Stillstand versetzt und die Warnblinkanlage aktiviert, um unter angemessener Beachtung der Verkehrssituation die größtmögliche Sicherheit für die Fahrzeuginsassen, andere Verkehrsteilnehmende und Dritte zu gewährleisten". In früheren Gesetzesentwürfen waren andere Definitionen vorgeschlagen wurden, die jedoch in der Literatur auf Kritik gestoßen waren.[56]

Der neue § 1e StVG regelt die Voraussetzungen des Betriebs von Fahrzeugen mit autonomer Fahrfunktion. Nach Abs. 1 ist der Betrieb eines solchen Kraftfahrzeugs zulässig, wenn das Kraftfahrzeug zum einen den technischen Voraussetzungen entspricht, die in § 1e Abs. 2 StVG ausgeführt werden, des Weiteren für das Kraftfahrzeug eine Betriebserlaubnis erteilt wurde, drittens das Kraftfahrzeug in einem von einer zuständigen Behörde genehmigten festgelegten Betriebsbereich eingesetzt wird und schließlich viertens das Kraftfahrzeug zur Teilnahme am öffentlichen Straßenverkehr zugelassen ist.

---

[54] Eingehend zu den „Stufen" des automatisierten und autonomen Fahrens *Herrmann/Brenner*, Die Autonome Revolution (Fn. 6), S. 59 ff.

[55] Zum Folgenden auch schon *Hilgendorf*, JZ 2021, 444 ff.

[56] *Hilgendorf*, JZ 2021, 444 (446).

2. Technische Vorgaben

In § 1e Abs. 2 StVG wird geregelt, über welche technische Ausrüstung Fahrzeuge mit autonomer Fahrfunktion in Deutschland verfügen müssen. Die Anforderungen sind durchaus anspruchsvoll:

a) Kernkompetenzen und „Leben-gegen-Leben-Entscheidungen"

Zunächst (in Nr. 1) wird sozusagen die Kernkompetenz autonomer Mobilität festgelegt: das Fahrzeug muss in der Lage sein, „die Fahraufgabe innerhalb des jeweiligen festgelegten Betriebsbereichs selbstständig zu bewältigen, ohne dass eine fahrzeugführende Person in die Steuerung eingreift oder die Fahrt des Kraftfahrzeugs permanent von der technischen Aufsicht überwacht wird". Wichtig ist also, dass das Fahrzeug seine Fahraufgaben selbstständig, ohne Eingriff eines Fahrzeugführers, erfüllen kann und zweitens, dass das Fahrzeug eigenständig unterwegs ist und von der technischen Aufsicht nicht permanent überwacht werden muss. Es handelt sich mithin nicht um teleoperiertes (ferngesteuertes), sondern um tatsächlich autonomes Fahren.[57]

Die rechtstheoretische und rechtsethisch interessanteste Bestimmung der Gesetzesreform findet sich in § 1e Abs. 2 Nr. 2. Dort wird festgelegt, dass autonome Fahrzeuge eine technische Ausrüstung haben müssen, die es ihnen ermöglicht, „selbstständig den an die Fahrzeugführung gerichteten Verkehrsvorschriften zu entsprechen". Sie müssen darüber hinaus ein „System der Unfallvermeidung" besitzen, welches (a) „auf Schadensvermeidung und Schadensreduzierung ausgelegt" ist, (b) „bei einer unvermeidbaren alternativen Schädigung unterschiedlicher Rechtsgüter die Bedeutung der Rechtsgüter berücksichtigt, wobei der Schutz menschlichen Lebens die höchste Priorität besitzt", und (c) „für den Fall einer unvermeidbaren alternativen Gefährdung von Menschenleben keine weitere Gewichtung anhand persönlicher Merkmale vorsieht". Der Gesetzgeber versucht hier, die Vorgaben der Ethikkommission zum automatisierten und vernetzten Fahren umzusetzen.[58]

Die Arbeitsergebnisse dieser Kommission können sich wiederum auf eine lange rechtsethische und rechtsphilosophische Debatte über Schadensvermeidung und „Leben gegen Leben- Konstellationen stützen:[59] Mit dem Grundsatz

---

[57] Wobei zu betonen ist, dass das Konzept der „Autonomie" in seinen philosophischen und psychologischen Dimensionen weitgehend ungeklärt ist. Unter (technischer) Autonomie soll hier und im Folgenden schlicht die Unabhängigkeit technischer Prozesse von einzelfallbezogenem menschlichen Input verstanden werden.

[58] Bericht der Ethik-Kommission für automatisiertes und vernetztes Fahren vom Juni 2017, S. 17 ff.; https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile.

[59] Dazu *Hilgendorf* ZStW 2018, 674 (883 ff.); BVerfG JZ 2022, 145 ff. mit Anm. *Hilgendorf*, ebenda, S. 153 ff.

der Schadensvermeidung und Schadensreduzierung wird ausgesagt, dass Schäden nach Möglichkeit zu vermeiden sind und, wenn eine Schadensherbeiführung unvermeidlich ist, der Schaden möglichst geringgehalten werden soll. Dieser allgemeine „Grundsatz des geringsten Übels" wird sodann für den Fall einer „unvermeidbaren alternativen Schädigung unterschiedlicher Rechtsgüter" näher ausgeführt. In solchen Fällen soll „die Bedeutung der Rechtsgüter berücksichtigt" werden. Dies dürfte so zu verstehen sein, dass auch in einer Dilemma-Situation, wenn also zwei oder mehr unterschiedliche Rechtsgüter in Gefahr sind, versucht werden muss, den Schaden so gering wie möglich zu halten. Eine solche Vorschrift setzt freilich einen Maßstab voraus, nach dem der Wert unterschiedlicher Rechtsgüter bemessen werden kann. Das Gesetz sagt dazu nichts, sondern legt lediglich fest, dass der Schutz menschlichen Lebens die „höchste Priorität" besitzen sollen, § 1e Abs. 2 Nr. 2b StVG.[60] Dies bedeutet, dass dann, wenn in einer Dilemma-Situation menschliches Leben mit anderen Rechtsgütern konkurriert, menschlichem Leben stets der Vorrang einzuräumen ist.

Im nächsten Satz gilt das Gesetz noch einen Schritt weiter und regelt den Sonderfall einer „unvermeidbaren alternativen Gefährdung von Menschenleben". Dies betrifft also den Fall, dass Menschenleben gegen Menschenleben steht. Auch diese Sachlage wird nicht vollständig normiert, der Gesetzgeber legt nur fest, dass in „Leben gegen Leben-" Konstellationen persönliche Merkmale keine Rolle spielen dürfen.[61] Dies sind Merkmale wie etwa das Geschlecht, das Alter[62] oder die Hautfarbe von Menschen. Keine Aussage trifft das Gesetz über Fragen der Quantität potentieller Opfer (also etwa wenn ein Leben gegen zwei oder mehr Leben steht). Nicht geregelt wird des Weiteren der Fall, in dem zwar Menschenleben unvermeidbar alternativ gefährdet sind, die Wahrscheinlichkeit einer Tötung aber unterschiedlich groß ist. Dies wäre etwa der Fall, wenn in einer Dilemma-Situation bei Wahl eines Ausweichweges x ein menschliches Leben mit sehr hoher Wahrscheinlichkeit vernichtet wird, bei Wahl des alternativen Ausweichweges y dagegen eine Tötung von Menschen zwar möglich, aber doch eher unwahrscheinlich erscheint.[63]

---

[60] Dass nach der Konzeption des Grundgesetzes der Schutz des Lebens, und nicht der der Würde, Art. 1 Abs. 1 GG, höchste Priorität genießt, lässt sich damit legitimieren, dass Würdeverletzungen kaum durch Bordsensoren festgestellt werden können. Außerdem dürften Würdeverletzungen im Straßenverkehr seltener vorkommen als Verletzungen der körperlichen Unversehrtheit und des Lebens.

[61] Dazu auch *Steiner*/*Steege*, ZdiW 2022, 1115 f.

[62] Es ist bemerkenswert, dass der Gesichtspunkt „Alter" in den Auseinandersetzungen um die Zulässigkeitsvoraussetzungen einer Triage in jüngerer Zeit mehrfach als relevant eingestuft wurde, so etwa von *Hoven*, JZ 2020, 449 (451).

[63] Zuverlässige quantifizierte Wahrscheinlichkeiten einer Schadensherbeiführung dürften nur selten zur Verfügung stehen, komparative Angaben (wie im Text) aber durchaus, *Hilgendorf*, ZStW 2018, 674 (697 ff.).

b) Weitere technische Anforderungen

Darüber hinaus werden in den Nrn. 3–10 weitere Anforderungen an die technische Ausrüstung von Kraftfahrzeugen mit autonomer Fahrfunktion formuliert. So muss der Bordcomputer in der Lage sein, „das Kraftfahrzeug selbstständig in einen risikominimalen Zustand zu versetzen, wenn die Fortsetzung der Fahrt nur durch eine Verletzung des Straßenverkehrsrechts möglich wäre." Damit sind etwa Fälle gemeint, in denen eine defekte Ampel permanent auf Rot steht und ein Weiterkommen nur möglich ist, wenn die rote Ampel durchfahren wird. In derartigen Fällen muss das Fahrzeug außerdem in der Lage sein, der technischen Aufsicht selbstständig Fahrmanöver vorzuschlagen, die eine Fortsetzung der Fahrt ermöglichen und darüber hinaus Daten liefern, die der technischen Aufsicht eine Beurteilung der Situation und damit auch eine Freigabe des vorgeschlagenen Fahrmanövers möglich machen (Nr. 4).

Das Gesetz legt des Weiteren fest, dass der Bordcomputer ein von der technischen Aufsicht vorgegebenes Fahrmanöver überprüfen können muss. In dem Fall, dass das Fahrmanöver eine am Verkehr teilnehmende oder unbeteiligte Personen gefährden würde, darf das Fahrmanöver nicht ausgeführt werden, vielmehr muss der Bordcomputer das Kraftfahrzeug eigenständig in einen risikominimalen Zustand versetzen (Nr. 5). Hierin drückt sich erneut die Vorstellung von menschlichem Leben als dem „Höchstwert" aus.[64]

Das Fahrzeug muss außerdem über eine technische Ausrüstung verfügen, die es möglich macht, „eine Beeinträchtigung ihrer Funktionalität der technischen Aufsicht unverzüglich anzuzeigen" (Nr. 6). Das Fahrzeug muss also in der Lage sein, Fehler im eigenen System zu entdecken und zu melden. Nach Nr. 7 muss das Fahrzeug außerdem die Fähigkeit besitzen, seine „Systemgrenzen zu erkennen und beim Erreichen einer Systemgrenze, beim Auftreten einer technischen Störung, die die Ausübung der autonomen Fahrfunktion beeinträchtigt, oder beim Erreichen der Grenzen des festgelegten Betriebsbereichs das Kraftfahrzeug selbstständig in einen risikominimalen Zustand zu versetzen". Auf diese Weise soll Überforderungen des Bordsystems vorgebeugt und die Gefahr für die Passagiere und andere Verkehrsteilnehmer minimiert werden.

Die technische Aufsicht, so legt § 1e Abs. 2 Nr. 8 StVG fest, muss jederzeit in der Lage sein, das Fahrzeug zu deaktivieren. Auch Fahrzeuginsassen müssen die Möglichkeit zur Deaktivierung des Fahrzeuges besitzen. Im Falle einer solchen Deaktivierung muss das Bordsystem das Kraftfahrzeug selbstständig in den risikominimalen Zustand versetzen.

---

[64] Wobei darauf hinzuweisen ist, dass das Rechtsgut „Leben" in Art. 2 Abs. 2 Satz 1 GG unter einem Gesetzesvorbehalt steht, während die Menschenwürde in Art. 1 Abs. 1 GG als „unantastbar" bezeichnet wird und jede Einschränkung ihres Schutzbereichs als rechtswidrig angesehen wird, dazu *Kunig/Kotzur*, Art. 1 Rn. 17, in: von Münch/Kunig (Hrsg.), Grundgesetzkommentar, Bd. 1, 7. Aufl. 2021.

Das Fahrzeug muss in der Lage sein, „der technischen Aufsicht das Erfordernis der Freischaltung eines alternativen Fahrmanövers, der Deaktivierung mit ausreichender Zeitreserve sowie Signale zum eigenen Funktionsstatus optisch, akustisch oder sonst wahrnehmbar anzuzeigen", § 1e Abs. 2 Nr. 9 StVG. Diese Bestimmung bezieht sich auf die Möglichkeit einer angemessenen Kommunikation zwischen Fahrzeug und technischer Aufsicht. Dieser Gesichtspunkt wird, aus einer anderen Perspektive, auch von § 1e Abs. 2 Nr. 10 StVG angesprochen. Danach muss das Fahrzeug über eine technische Ausrüstung verfügen, welche „ausreichend stabile und vor unautorisierten Eingriffen geschützte Funkverbindungen, insbesondere zur Technischen Aufsicht" sicherstellen kann. Angesichts der derzeit noch bestehenden Probleme bei der flächendeckenden Funkverbindung in Deutschland erscheint diese Anforderung als sehr ambitioniert. Hinzu kommt die immer weiter wachsende Gefahr von Cyberangriffen.[65] Schließlich wird in Nr. 10 festgelegt, dass das Kraftfahrzeug selbstständig in einen risikominimalen Zustand versetzt werden muss, wenn die Funkverbindung abbricht oder unerlaubt darauf zugegriffen wird.

c) Zum Einsatz lernfähiger KI

Lernfähige KI ist im geltenden Straßenverkehrsrecht autonomer Systeme (§§ 1a–1l StVG) nicht geregelt. In einer nicht Gesetz gewordenen Entwurfsfassung war vorgesehen, solche Systeme zuzulassen, die Umsetzung erlernter Verhaltensweisen durch autonom fahrende Kraftfahrzeuge aber von einer vorherigen Genehmigung des Kraftfahrtbundesamts abhängig zu machen.[66] Dieser Ansatz erscheint gerade im Hinblick auf einen wirksamen Verbraucherschutz zukunftsfähig, setzt aber voraus, dass das Kraftfahrtbundesamt über die nötigen Ressourcen und Fachkompetenzen verfügt, um derartige Überprüfungen durchführen zu können. Dabei sollte eine enge Zusammenarbeit mit der Automobilindustrie angestrebt werden.

3. Zur Pflichtenstellung der Beteiligten

§ 1f StVG legt die Pflichten der Beteiligten beim Betrieb von Kraftfahrzeugen mit autonomer Fahrfunktion fest. Dabei werden die Pflichten des Halters, der technischen Aufsicht und der Hersteller eines Kraftfahrzeugs mit autonomer Fahrfunktion unterschieden.

Die Verpflichtung zur Erhaltung der Verkehrssicherheit und der Umweltverträglichkeit eines Fahrzeugs liegt beim Halter, § 1f Abs. 1 StVG. Damit wird die traditionelle Halterorientierung des deutschen Straßenverkehrsrechts fortge-

[65] Aktuelle Informationen dazu auf den Internetseiten des Bundesamts für Sicherheit in der Informationstechnik, www.bsi.de.

[66] Näher *Hilgendorf*, JZ 2021, 444 (449 f.).

schrieben.[67] Er hat die dafür erforderlichen Vorkehrungen zu treffen. Dies bedeutet insbesondere die regelmäßige Wartung der für die autonome Fahrfunktion erforderlichen Systeme. Der Halter hat außerdem dafür zu sorgen, „dass die sonstigen, nicht an die Fahrzeugführung gerichteten Verkehrsvorschriften eingehalten werden". Das betrifft insbesondere solche Vorschriften, die sich auf die Passagiere beziehen, also etwa Anschnallvorschriften.[68] Schließlich, und dies verdient besondere Hervorhebung, liegt es in der Zuständigkeit des Halters, sicherzustellen, dass die Aufgaben der technischen Aufsicht erfüllt werden (Nr. 3). In der Regel wird der Halter dafür eine andere Person einzustellen haben.

Außerdem werden (in § 1f Abs. 2 StVG) die Aufgaben der Technischen Aufsicht festgelegt. Dazu gehören die Freischaltung alternativer Fahrmanöver (Nr. 1) die Deaktivierung in Gefahrfällen (Nr. 2), die Bewertung von Signalen der technischen Ausrüstung zum eigenen Funktionsstatus (Nr. 3) sowie die Aufnahme von Kontakt zu den Insassen des autonomen Fahrzeugs und die Verkehrssicherung, wenn das Kraftfahrzeug in den risikominimalen Zustand versetzt wird (Nr. 4).

Besonders ausführlich werden die Pflichten des Herstellers von Kraftfahrzeugen mit autonomer Fahrfunktion geregelt. Über den gesamten Entwicklungs- und Betriebszeitraum des Fahrzeugs hinweg muss gegenüber dem Kraftfahrtbundesamt (KBA) gewährleistet werden, dass das Fahrzeug gegen Cyberangriffe geschützt ist. Nötig ist außerdem eine Risikobeurteilung des Kraftfahrzeugs sowie der Nachweis einer ausreichend sicheren Funkverbindung (Nr. 3). Hinzu tritt das Erfordernis einer ausführlichen Systembeschreibung. (Nr. 4). Bemerkenswert und über das bisherige Produkthaftungsrecht hinausgehend ist die in Nr. 5 festgelegte Verpflichtung der Hersteller, Schulungen für die am Betrieb autonomer Fahrzeuge beteiligten Personen anzubieten. Darin soll die technische Funktionsweise der Fahrzeuge „insbesondere im Hinblick auf die Fahrfunktion und die Aufgabenwahrnehmung der technischen Aufsicht" dargelegt werden. Falls Manipulationen am Fahrzeug oder dessen elektronischer oder elektrischer Architektur erkannt werden, sollen unverzüglich die zuständigen Behörden informiert werden (Nr. 6).[69] Der Hersteller unterliegt somit umfangreichen, im Gesetz detailliert beschrieben Überwachungs- und Sicherungspflichten.

Der Halter muss darüber hinaus beim Betrieb des Fahrzeugs bestimmte Daten speichern. Details dazu sind in § 1g StVG festgelegt. Zu speichern sind danach Fahrzeugidentifizierungsnummer, Positionsdaten, Anzahl und Zeiten der Nutzung sowie der Aktivierung und der Deaktivierung der autonomen Fahrfunktion, Anzahl und Zeiten der Freigabe von alternativen Fahrmanövern, Systemüberwachungsdaten einschließlich Daten zum Softwarestand, Umwelt und Wetter-

---

[67] Dazu eingehend *Greger/Zwickel*, Haftung im Straßenverkehr, 6. Aufl. 2021, § 3.

[68] Vgl. §§ 21, 21a StV0.

[69] Dies dürfte ausreichen, um von einer „Garantenstellung" i. S. d. Strafrechts sprechen zu können.

bedingungen, Vernetzungsparameter wie beispielsweise Übertragungslatenz und verfügbare Bandbreite, der Name der aktivierten und deaktivierten passiven und aktiven Sicherheitssysteme, Daten zum Zustand dieser Sicherheitssysteme sowie die Instanz, die das Sicherheitssystem ausgelöst hat, Daten zur Fahrzeugbeschleunigung in Längs- und Querrichtung, Geschwindigkeit, der Status der lichttechnischen Einrichtungen, die Spannungsversorgung des Kraftfahrzeugs mit autonomer Fahrfunktion und von extern an das Kraftfahrzeug gesendete Befehle und Informationen. Diese Daten muss der Halter unter Umständen auch dem Kraftfahrtbundesamt und anderen zuständigen Behörden übermitteln können. Eine dauerhafte Speicherung ist vorgesehen bei Eingriffen durch die technische Aufsicht, bei Konfliktszenarien, insbesondere bei Unfällen und Fast-Unfall-Szenarien,[70] bei nicht planmäßigem Spurwechsel oder Ausweichen sowie bei Störungen im Betriebsablauf.

Einzelheiten der Zulassung des Betriebs von autonomen Fahrzeugen auf öffentlichen Straßen soll eine Rechtsverordnung regeln, die vom Bundesministerium für Verkehr und digitale Infrastruktur erlassen werden kann, § 1j StVG.[71] Das Gesetz soll nach Ablauf des Jahres 2023 insbesondere im Hinblick auf die Auswirkungen auf die Entwicklung des autonomen Fahrens, und die Vereinbarkeit mit Datenschutzvorschriften evaluiert werden, § 1l StVG.

4. Vorläufige Bewertung

Überblickt man das neue Gesetz im Gesamtzusammenhang, so wird deutlich, dass der Gesetzgeber versucht, die Interessen der Halter, der Hersteller, der Passagiere autonomer Fahrzeuge und sonstiger Verkehrsteilnehmer so auszutarieren, dass ein nachhaltiger, gemeinwohlorientierter Kompromiss zwischen Mobilität und Sicherheitsinteressen einerseits, praktischer Implementierbarkeit und ökonomischem Betrieb andererseits erreicht werden kann. Das autonome Fahren kommt vielen Belangen moderner Mobilitätsplanung entgegen. Es kann jedoch nur dann auf Akzeptanz hoffen, wenn Gefahren durch Fehlfunktionen des Fahrzeugs selbst, Probleme der Interaktion mit anderen, möglicherweise nicht hoch technisierten Fahrzeugen und schließlich auch Probleme mit der Cybersicherheit angemessen adressiert werden.

Das neue Gesetz schafft dafür einen sinnvollen und in sich schlüssigen Rahmen. Die Hersteller und Halter werden aufgefordert, technische Lösungen für die angesprochenen Probleme zu finden. Diese Aufgabe ist alles andere als einfach und stellt höchste Anforderungen an Erfindungsreichtum und den Ingenieurs-

---

[70] Ein Problem besteht darin, wie ein „Fast-Unfall-Szenario" begrifflich zu bestimmen ist. Bei einem weiten Begriffsverständnis ließe sich über die im Text genannte Bestimmung auch die Aufzeichnung des „normalen" Straßenverkehrs legitimieren.

[71] Die Verordnung soll Zeitungsberichten zufolge im März 2022 vorliegen.

sachverstand. Besonders hervorzuheben ist, dass die meisten der skizzierten Anforderungen Softwarelösungen erfordern, ein Gebiet, auf welchem die deutsche Automobilindustrie bislang nicht unbedingt zu den Spitzenreitern zählt. Es bleibt abzuwarten, welchen Herstellern es gelingen wird, die ersten entsprechenden Fahrzeuge anzubieten.

Infolge des im Entwurf der KI-VO verwendeten sehr weiten KI-Begriffs fallen, wenn der Entwurf nicht noch geändert wird, auch digitale Systeme in Fahrzeugen als Hochrisiko-KI in den Anwendungsbereich der KI-VO (§ 6). Was dies im Detail für die Hersteller und Nutzer von Fahrzeugen bedeutet, wird bald im Detail zu klären sein. Insbesondere gilt es zu klären, ob bzw. inwieweit die KI-VO in ihrer endgültigen Fassung eine Reform des deutschen Straßenverkehrsrechts erforderlich machen wird.

VI. Zusammenfassung und Ausblick

KI dringt derzeit rasch in alle Bereiche von Staat und Gesellschaft vor und macht auch vor der Mobilität nicht Halt. Rein bereichsspezifische Regulierungen verlieren deshalb an Attraktivität. Hinzu kommt, dass gesamteuropäische Lösungsansätze bloß nationalen Regelungsversuchen vorzuziehen sind. Es verdient deshalb Zustimmung, dass die EU-Kommission auf Grundlage der Vorarbeiten der HLEG AI einen Regelungsvorschlag für den Umgang mit KI ausgearbeitet und vorgestellt hat.[72] Darin werden auch Fragen KI-gestützter Mobilität geregelt.[73] Der Vorschlag, der auf die Verabschiedung einer EU-Verordnung abzielt, schießt allerdings in Teilen über das Ziel einer Rahmensetzung hinaus, ist übermäßig restriktiv und lässt vor allem die erforderliche Rechtssicherheit für die Unternehmen vermissen. Im Hinblick auf den Schutz der Verbraucherinnen und Verbraucher knüpft der Entwurf an die älteren europäischen Vorgaben an und führt sie fort.

Das deutsche Gesetz, mit dem im Sommer 2021 das autonome Fahren auf Deutschlands Straßen in vordefinierten Betriebsbereichen zugelassen wurde,[74] stellt die erste Regelung autonomen Fahrens im Regelbetrieb weltweit dar. Die Reform scheint gelungen; in einigen seiner Bestimmungen, etwa zum technischen Umgang mit Dilemma-Situationen, bewegt sie sich allerdings am Rand des rechtstechnisch Darstellbaren. Im Hinblick auf die Belange der Verbraucherpolitik orientiert sie sich an deutschen und europäischen Vorgaben; besonders hervorzugeben ist, dass das bewährte Haftungsmodell: Gefährdungshaftung des Halters plus Pflichtversicherung – beibehalten und auf die Person der technischen Aufsicht ausgeweitet wurde.

---

[72] Siehe oben III, IV.
[73] Siehe oben IV.2 und 3.
[74] Siehe oben V.

Dennoch ist die Verbraucherpolitik im Bereich KI-gestützter Mobilität mit neuen Herausforderungen konfrontiert: Angesichts der Schwäche deutscher Softwareunternehmen und vielfältiger Übertreibungen im Datenschutz[75] droht auch im Bereich der Mobilität eine Abhängigkeit von US-Monopolen, die zum Einsickern neuer Geschäftsmodelle und damit zur Verbreitung anderer, häufig deutlich weniger verbraucherfreundlicher Vorstellungen von Risikoverteilung und Verbraucherpolitik führen könnte. Nicht jeder Wandel ist ein Wandel zum Besseren. Wer die europäischen Standards in der Verbraucherpolitik verteidigen und stärken will, darf sich nicht von außereuropäischen Großunternehmen abhängig machen.[76] Klärungsbedarf besteht im Hinblick auf die Vorgaben der derzeit vorbereiteten EU-KI-Verordnung. Nach jetzigem Stand dürften KI-Systeme in Kraftfahrzeugen als „Hochrisiko-KI" im Sinne der Verordnung einzustufen sein. Was dies im Detail für die Kraftfahrzeughersteller und Verbraucher bedeutet, muss bald herausgearbeitet werden.

[75] Wobei es nicht die gesetzlichen Normierungen selbst sind, die Probleme bereiten, sondern ein unreflektierter, juristisch oft gänzlich unaufgeklärter „Datenschutzabsolutismus", der im Normenbestand, insbesondere in der DSGVO, keine Stütze findet.

[76] Damit einher geht dann meist auch eine Schwächung der eigenen Staatlichkeit, dazu *Schallbruch*, Schwacher Staat im Netz. Wie die Digitalisierung den Staat in Frage stellt, 2018.

Trainingsdaten und die Rechte von betroffenen Personen

– in der DSGVO und darüber hinaus?

Gerrit Hornung[1]

I. Einleitung

Die Herausforderungen algorithmenbasierter Datenverarbeitung werden in den Rechtswissenschaften seit etlichen Jahren diskutiert. Dies begann bereits in einer Zeit, in der es noch um die rechtlichen Anforderungen an die Verwendung herkömmlicher Algorithmen ging, die also noch nicht als Künstliche Intelligenz (KI) bezeichnet werden konnten.[2] Mit der fortschreitenden Entwicklung der Modelle und Algorithmen, etwa im Bereich von Technologien maschinellen Lernens,[3] intensivierten sich die entsprechenden rechtswissenschaftlichen Analysen.[4]

Erst in jüngerer Zeit gewinnt demgegenüber die Erkenntnis Raum, dass es sich bei den identifizierten rechtlichen Problemen der KI teilweise nicht um solche der verwendeten Algorithmen selbst handelt. Während prinzipielle Probleme

---

[1] Der Text ist im Zusammenhang mit Arbeiten des BMBF-Projekts „Künstliche Intelligenz zur Analyse und Fusion von Erdbeobachtungs- und Internetdaten zur Unterstützung bei der Lageerfassung und -einschätzung“ (AIFER, FKZ 13N15528) sowie der Projektgruppe „Nachhaltige Intelligenz – intelligente Nachhaltigkeit“ des hessischen Zentrums verantwortungsbewusste Digitalisierung (ZEVEDI) entstanden, deren Sprecher der Verfasser ist.

[2] S. z. B. die Diskussion um die (In-)Transparenz der Scorewert-Berechnung durch die SCHUFA; dazu BGHZ 200, 38; *Hammersen/Schade*, DuD 2014, 399; *Paal*, JZ 2014, 1006.

[3] Insbesondere Verfahren des deep learning, dazu aus technischer Sicht *Goodfellow/Bengio/Courville*, Deep Learning, 2016; *Russell/Norvig*, Artificial Intelligence: A Modern Approach, 4. Auflage 2020, 750 ff.; zu den soziotechnischen Voraussetzungen des deep learning s. *Mühlhoff*, ZfM 2/2019, 56.

[4] Grundsätzlich zu den rechtlichen Herausforderungen der KI z. B. *Hoffmann-Riem*, AöR 142 (2017), 1; *Wischmeyer*, AöR 143 (2018), 1; *Herberger*, NJW 2018, 2825; *Martini*, Blackbox Algorithmus, 2019; s. ferner die Beiträge in *Kaulartz/Braegelmann* (Hrsg.) Rechtshandbuch AI and Machine Learning, 2020; *Beck/Kusche/Valerius* (Hrsg.), Digitalisierung, Automatisierung, KI und Recht, 2020; *Ebers/Heinze/Krügel/Steinrötter* (Hrsg.), Künstliche Intelligenz und Robotik, 2020; *Wischmeyer/Rademacher* (Hrsg.), Regulating Artificial Intelligence, 2020; *Chibanguza/Kuß/Steege* (Hrsg.), Künstliche Intelligenz. Recht und Praxis automatisierter und autonomer Systeme, 2021; *Dederer/Yu-Cheol Shin* (Hrsg.), Künstliche Intelligenz und juristische Herausforderungen, 2021; *Ebers/Steinrötter* (Hrsg.), Künstliche Intelligenz und smarte Robotik im IT-Sicherheitsrecht, 2021; *Busch/De Franceschi* (Hrsg.), Algorithmic Regulation and Personalized Law, 2021; zu den Herausforderungen für die demokratische Gesellschaft s. die Beiträge in *Unger/v. Ungern-Sternberg*, Demokratie und künstliche Intelligenz, 2019; s. a. das Gutachten der *Datenethikkommission*, 2019, 159 ff.

der Lesbarkeit der Entscheidungen und ihrer – hiervon zu unterscheidenden – Verständlichkeit bzw. Nachvollziehbarkeit für Menschen eher vom verwendeten KI-Modell abhängen (also z. B. ein bestimmter Subtyp neuronaler Netze), werden Verzerrungseffekte oftmals durch das Training dieses Modells hervorgerufen. Dabei werden Trainingsdaten verwendet und im Rahmen des Trainings die Parameter des Modells kontinuierlich angepasst, um die Ergebnisgenauigkeit zu verbessern. Dementsprechend können Fehler entweder dadurch hervorgerufen werden, dass die zum Training der Algorithmen verwendeten Daten fehlerhaft, nicht repräsentativ oder in anderer Weise ungeeignet sind, oder durch eine fehlerhafte Anpassung der Parameter. Datenqualität und Parametrierung hängen insoweit zusammen, als z. B. eine gewisse fehlende Repräsentativität der Trainingsdaten (wenn diese beispielsweise deutlich mehr Daten von Männern als von Frauen enthalten) durch entsprechende Parameter ausgeglichen werden kann, wenn sie erkannt wird.

Fehlt es an einer in dieser Weise notwendigen Reaktion auf verzerrende Trainingsdaten oder weisen diese prinzipielle Beschränkungen auf, die durch die Parametrierung nicht ausgeglichen werden können, so besteht ein erhebliches Risiko, dass das spezifische Zusammenwirken des verwendeten Algorithmus mit den ihm zur Verfügung gestellten Trainingsdaten zu Verzerrungs- und Diskriminierungseffekten führt.[5] Dies wirft etliche Rechtsprobleme auf, deren Diskussion noch in ihren Anfängen steht.[6] Zwei wichtige Fragen sind dabei die nach den bereits geltenden rechtlichen Anforderungen an den Umgang mit Trainingsdaten einerseits, der Sinnhaftigkeit und etwaigen Ausgestaltung einer entsprechenden Regulierung andererseits.

Da viele Trainingsdaten personenbezogen sind oder ein Personenbezug zumindest nicht ausgeschlossen werden kann, liegt es nahe, beide Fragen (auch) aus

---

[5] Zu den Diskriminierungsrisiken des Einsatzes von KI z. B. *Ernst*, JZ 2017, 1026, 1032 ff.; *Wischmeyer*, AöR 143 (2018), 1, 26 ff.; *Steege*, MMR 2019, 715; *Wildhaber/Lohmann/Kasper*, ZSchwR I 2019, 459; *Mann/Matzner*, Big Data & Society 2019, 1; *Martini*, Blackbox Algorithmus, 2019, 47 ff., 73 ff.; *Wildhaber/Lohmann/Kasper*, ZSchwR I 2019, 459; *Mann/Matzner*, Big Data & Society 2019, 1; *Beck/Grunwald/Jacob/Matzner*, Künstliche Intelligenz und Diskriminierung. Whitepaper der Plattform Lernende Systeme, 2019; *Kolleck/Orwat*, Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick. TAB-Hintergrundpapier Nr. 24, 2020; *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2020; *Hacker*, ZGE 2020, 239, 251 ff.; s. a. die Unterrichtung der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, BT-Drs. 19/23700, 60 ff. sowie Entschließung des Europaparlaments v. 6.10.2021, P9_TA(2021)0405, in der die Risiken von Verzerrungen in Trainingsdaten betont werden (Rn. 8) und ihre Dokumentation gefordert wird (Rn. 19).

[6] S. bisher v. a. *Hacker*, ZGE 2020, 239; *ders.*, GRUR 2020, 1025 sowie den Beitrag in diesem Band; s. ferner aus v. a. datenschutzrechtlicher Perspektive *Heinemeyer*, CR 2019, 147; *Niemann/Kevekordes*, CR 2020, 17; *dies.*, CR 2020, 179; *Raji*, DuD 2021, 303; *Boenisch*, DuD 2021, 448; *Valkanova*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.1; *Kaulartz*, ebd., Kap. 8.9; *Vogel*, Künstliche Intelligenz und Datenschutz, 2022, 49 ff., 69 f.; zur Haftung für Trainingsdaten *Hacker*, ZGE 2020, 239, 249 ff.; *Zech*, NJW 2022, 502.

datenschutzrechtlicher Sicht zu analysieren. Dies kann wiederum aus einer eher objektivrechtlichen Perspektive erfolgen, die Datenschutz maßgeblich als Instrument zur Begrenzung von Datenmacht begreift,[7] oder aus der Perspektive der betroffenen Person im Sinne von Art. 4 Nr. 1 DSGVO, bei der es sich typischerweise um eine Verbraucherin oder einen Verbraucher handeln wird. Der Beitrag wählt die zweite Perspektive und untersucht – im Sinne des Leitthemas der Verbraucherrechtstage 2021 –, welchen Beitrag das Datenschutzrecht leisten kann, um eine vertrauenswürdige Verwendung von KI in Deutschland und Europa zu befördern. Der Text erläutert zunächst Interessenlagen und Herausforderungen, bevor die datenschutzrechtlichen Betroffenenrechte auf ihre Leistungsfähigkeit zum Schutz der betroffenen Person vor unangemessener und zu weitreichender Datenverarbeitung befragt werden. Im letzten Schritt wird diskutiert, ob die Unzulänglichkeiten des gerade datenschutzrechtlichen Zugriffs auf das Problem der Trainingsdaten regulatorisch angegangen werden sollten, und ob der Vorschlag der europäischen Kommission für ein „Gesetz über Künstliche Intelligenz"[8] (oftmals auch „KI-Verordnung" genannt; im Folgenden KOM-E) insoweit Verbesserungen bringen würde.

Noch während des europäischen Gesetzgebungsverfahrens hat sich der schleswig-holsteinische Gesetzgeber entschlossen, mit dem seit dem 15.4.2022 geltenden § 8 des schleswig-holsteinischen IT-Einsatz-Gesetz (ITEG SH)[9] eine erste verbindliche nationale Regulierung zu Trainingsdaten zu verabschieden.[10] Auch wenn eine verabschiedete KI-Verordnung diese Regelungen verdrängen bzw. modifizieren wird, ist dieser erste Schritt bemerkenswert; die Norm wird deshalb im Folgenden an geeigneten Stellen berücksichtigt.

---

[7] S. zu diesem Ansatz z. B. *v. Lewinski*, Die Matrix des Datenschutzes, 2014, 55 ff.; in historischer Perspektive *v. Lewinski*, in: Arndt u. a. (Hrsg.), Freiheit – Sicherheit – Öffentlichkeit. 48. Assistententagung Öffentliches Recht, 2009, 201 ff.

[8] 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 v. 21.4.2021, COM(2021) 206 final; zu Hintergründen und ersten Analysen s. z. B. *Veale/Borgesius*, CRi 2021, 97; *Geminn*, ZD 2021, 354; *Spindler*, CR 2021, 361; *Grützmacher*, CR 2021, 433; *Kau*, CR 2021, 498; *Kalbhenn*, ZUM 2021, 663; *Valta/Vasel*, ZRP 2021, 142; *Ebers*, RDi 2021, 588; *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188; *Rostalski/Weiss*, ZfDR 2021, 329; *Ebers/Hoch/Rosenkranz/Ruschemeier/Steinrötter*, RDi 2021, 528; *Wiebe*, BB 2022, 899; zur Entwicklung *Hacker*, NJW 2020, 2142; breitere Analyse des Entwurfs mit Blick auf eine Regulierung von KI bei *Kau*, ZG 2021, 217; zu verbleibenden Regulierungsspielräumen der Mitgliedstaaten *Hornung*, DuD 2022, 561.

[9] Gesetz über die Möglichkeit des Einsatzes von datengetriebenen Informationstechnologien bei öffentlich-rechtlicher Verwaltungstätigkeit (IT-Einsatz-Gesetz – ITEG), GVBl. SH Nr. 5, 296; s. die Begründung, LT-Drs. SH 19/3267, 15 f., 69 f., 136 ff.

[10] Ein solches „Vorpreschen" nationaler Gesetzgeber lässt sich auch in anderen Regelungsbereichen mitunter beobachten, beispielsweise bei der IT-Sicherheitsregulierung (IT-Sicherheitsgesetz und NIS-Richtlinie) oder der Plattformregulierung (Netzwerkdurchsetzungsgesetz und Digital Services Act).

II. „Gute" Trainingsdaten

Ob Verbraucherschutz durch „gute" Trainingsdaten befördert werden kann, hängt zunächst von der Frage ab, was gute und was schlechte Trainingsdaten sind. Dies wird vielfach von den spezifischen Anforderungen zunächst des Trainings- und später des Einsatzszenarios des KI-Algorithmus abhängen. Verallgemeinernd lassen sich generische Anforderungen wie Repräsentativität und Aktualität festmachen.[11] Jedenfalls in erster Näherung bieten auch Art. 10 Abs. 3 und Abs. 4 KOM-E – sowie nunmehr als erste nationale Regelung § 8 Abs. 3 ITEG SH – sinnvolle Anhaltspunkte für die Frage, was qualitativ hochwertige Trainingsdaten ausmacht.

1. Begriff

Die Europäische Kommission schlägt in Art. 10 Abs. 3 S. 1 KOM-E vor, Trainings-, Validierungs- und Testdatensätze einheitlichen Anforderungen zu unterwerfen, soweit es sich um Hochrisiko-KI-Systeme (Art. 6 KOM-E)[12] handelt. Nach den Begriffsbestimmungen des KOM-E sind

- „Trainingsdaten" Daten, die zum Trainieren eines KI-Systems verwendet werden, wobei dessen lernbare Parameter und die Gewichte eines neuronalen Netzes angepasst werden (Art. 3 Nr. 29 KOM-E),
- „Validierungsdaten" Daten, die zum Bewerten des trainierten KI-Systems und zum Abstimmen seiner nicht lernbaren Parameter und seines Lernprozesses verwendet werden, um unter anderem eine Überanpassung zu vermeiden; der Validierungsdatensatz kann ein separater Datensatz oder Teil des Trainingsdatensatzes mit fester oder variabler Aufteilung sein (Art. 3 Nr. 30 KOM-E), und
- „Testdaten" Daten, die für eine unabhängige Bewertung des trainierten und validierten KI-Systems verwendet werden, um die erwartete Leistung dieses Systems vor dessen Inverkehrbringen oder Inbetriebnahme zu bestätigen (Art. 3 Nr. 31 KOM-E).

Diese Definitionen entsprechen einem üblichen Vorgehensmodell beim Training von KI-Systemen.[13] Für dieses Training steht in der Praxis eine gewisse Menge

---

[11] S. den Beitrag von *Abedjan* in diesem Band; s. ferner das Gutachten der *Datenethikkommission*, 2019, 94, 168 sowie aus rechtlicher Sicht näher *Hacker*, ZGE 2020, 239, 262 ff.

[12] Der KOM-E unterscheidet in einem risikobasierten Ansatz zwischen verbotenen Praktiken im Bereich der KI (Art. 5 KOM-E), Anforderungen an Hochrisiko-KI-Systeme (v. a. Art. 6–15 KOM-E) und sonstigen KI-Systemen; s. näher *Geminn*, ZD 2021, 354, 355 ff.; *Spindler*, CR 2021, 361, 362; *Valta/Vasel*, ZRP 2021, 142 f.; *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188, 1189 ff.; *Bomhard/Merkle*, RDi 2021, 276, 279 ff.; *Rostalski/Weiss*, ZfDR 2021, 329, 337 ff.; allgemein zu datenschutzrechtlichen Risikokriterien für KI *Rost*, DuD 2018, 558, 561 ff.

[13] *Niederée/Nejdl*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik, Rechtshandbuch, 2020, § 2 Rn. 28 ff.; § 8 Abs. 1 ITEG SH spricht von Daten „zum Zweck der Entwicklung und des Trainings", ohne beide Begriffe allerdings zu definieren.

von Daten zur Verfügung, auf die das System angewendet werden kann (z. B. um ein Objekt in einem Bild zu identifizieren). Diese Menge wird in zwei Teilmengen aufgespalten: zum einen der Datensatz, der (als Trainingsdaten i. S. v. Art. 3 Nr. 28 KOM-E) zum initialen Training des Modells verwendet wird, zum anderen der – meist kleinere – Datensatz, mit dem (als Testdaten i. S. v. Art. 3 Nr. 31 KOM-E) nach Abschluss des Trainings geprüft wird, wie gut das Training funktioniert hat. Um dies beurteilen zu können, darf das trainierte System zuvor nicht mit den Testdaten in Berührung kommen. Die in Art. 3 Nr. 30 KOM-E geregelten Validierungsdaten sind ein Spezialfall der Trainingsdaten, mit denen das KI-System bewertet wird und seine nicht lernbaren Parameter sowie der Lernprozess abgestimmt werden. Dies bezieht sich insbesondere auf das Risiko einer Überanpassung („Overfitting"), bei der das System so stark auf die Trainingsdaten angepasst wurde, dass es auf anderen Daten nicht oder nicht in hinreichender Qualität funktioniert.

Im Ergebnis ist der Unterschied zwischen den drei Datentypen also eine reine Willensentscheidung desjenigen, der das System trainiert, validiert und testet – die Daten gehören nicht als solche in eine der drei Gruppen. Da alle drei Phasen zu einem Begriff des Trainings im weiteren Sinne gerechnet werden können (wenn das abschließende Testen nicht zufriedenstellend ausfällt, wird erneut trainiert; d. h. das Testen ist Teil des rekursiven Trainingszyklus), wird im Folgenden der Begriff der Trainingsdaten als Oberbegriff für die drei genannten Gruppen verwendet.[14] Soweit konkrete Bestimmungen des KOM-E thematisiert werden, folgt die Terminologie allerdings den Begriffsbestimmungen in Art. 3 Nr. 29–31 KOM-E.

Nach Art. 10 Abs. 3 S. 1 KOM-E müssen Trainings-, Validierungs- und Testdatensätze relevant, repräsentativ, fehlerfrei und vollständig sein. § 8 Abs. 3 ITEG SH formuliert völlig anders und verlangt, dass die bei der Entwicklung und beim Training – außerdem, über Art. 30 Abs. 3 KOM-E hinausgehend, auch beim Einsatz – der Systeme verwendeten Daten „nicht-diskriminierend, integer, objektiv und valide" sind. Diese völlige Überlappungsfreiheit der verwendeten Begriffe ist angesichts der Tatsache bemerkenswert, dass der KOM-E ein halbes Jahr vor dem Gesetzesentwurf zum ITEG SH veröffentlicht wurde; auch die Begründung nimmt (weder zu dieser Norm[15] noch an einer anderen Stelle überhaupt) auf den KOM-E Bezug. Wie groß die inhaltlichen Unterschiede der Begriffe sind, muss angesichts der Tatsache offenbleiben, dass weder die schleswig-holsteinische Begründung noch der KOM-E[16] Erläuterungen bieten.

Gemäß Art. 10 Abs. 3 S. 2 KOM-E haben Trainings-, Validierungs- und Testdatensätze geeignete statistische Merkmale aufzuweisen, gegebenenfalls auch

---

[14] S. zum Begriff auch *Niederée/Nejdl*, in: Ebers/Heinze/Krügel/Steinrötter (Fn. 13), § 2 Rn. 31.
[15] LT-Drs. SH 19/3267, 155.
[16] EG 44 S. 3 KOM-E wiederholt insoweit lediglich Art. 10 Abs. 3 S. 1 KOM-E.

bezüglich der Personen oder Personengruppen, auf die das Hochrisiko-KI-System bestimmungsgemäß angewandt werden soll. Diese Merkmale der Datensätze können durch einzelne Datensätze oder eine Kombination solcher Datensätze erfüllt werden (Art. 10 Abs. 3 S. 3 KOM-E). Ergänzt wird dies in Art. 10 Abs. 4 KOM-E durch Anforderungen mit Bezug auf die spätere Einsatzumgebung. Die Trainings-, Validierungs- und Testdatensätze müssen danach, soweit dies für die Zweckbestimmung erforderlich ist, den Merkmalen oder Elementen entsprechen, die für die besonderen geografischen, verhaltensbezogenen oder funktionalen Rahmenbedingungen, unter denen das Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll, typisch sind.

2. Interessenlagen

In etlichen Situationen werden sowohl der datenschutzrechtlich Verantwortliche (also typischerweise der Nutzer des KI-Systems)[17] als auch die betroffenen Personen ein Interesse an der Einhaltung dieser Vorgaben, also an der Verwendung qualitativ hochwertiger Trainingsdaten haben. Irrelevante, nicht repräsentative, fehlerhafte oder unvollständige Trainingsdaten zu verwenden, kann zu – im weitesten Sinne – falschen oder unangemessenen Ergebnissen führen.[18] Hieran haben typischerweise weder der Nutzer des KI-Systems noch derjenige ein Interesse, der Adressat einer KI-basierten Entscheidung wird.[19] Allerdings sind die Folgen derartiger falscher oder unangemessener Ergebnisse vielfach ungleich verteilt. Aus Sicht des Nutzers kann es sich um ein prinzipielles „GIGO"-Problem („Garbage In, Garbage Out") handeln, das zur Unbrauchbarkeit des gesamten Systems führt. Wenn demgegenüber falsche und unangemessene Ergebnisse lediglich in Einzelfällen auftreten, mag es aus Perspektive des Nutzers immer noch sinnvoll sein, das System zu verwenden, weil es beispielsweise kostengünstiger als die manuelle Bearbeitung ist oder im Durchschnitt bessere Ergebnisse liefert.[20] Den im konkreten Einzelfall negativ Betroffenen, beispielsweise diskriminierten Verbraucherinnen und Verbrauchern, wird eine solche Effizienzbetrachtung hingegen nicht gerecht.

Aus einer datenschutzrechtlichen Perspektive sollte außerdem nicht übersehen werden, dass auch der Einsatz besonders guter Trainingsdaten ein Problem für

---

[17] Hier verwendet i. S. v. Art. 3 Nr. 4 KOM-E: eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System in eigener Verantwortung verwendet, es sei denn, das KI-System wird im Rahmen einer persönlichen und nicht beruflichen Tätigkeit verwendet.

[18] S. zu den Diskriminierungsrisiken die Nachweise in Fn. 5; zu Haftungsfragen *Hacker*, ZGE 2020, 239, 249 ff.; *Zech*, NJW 2022, 502.

[19] Zu den Interessen und den mit ihnen verbundenen „regulatorischen Risiken" s. *Hacker*, ZGE 2020, 239, 243 ff.

[20] Insoweit sind Anforderungen an die Qualität von Trainingsdaten wie in Art. 10 Abs. 5 KOM-E nur auf den ersten Blick „eigentlich selbstverständliche Anforderungen" (so *Spindler*, CR 2021, 361, 367).

Verbraucherinnen und Verbraucher sein kann. Qualitativ hochwertige KI-basierte Entscheidungen können zu ihren Gunsten, sehr leicht aber auch zu ihren Lasten eingesetzt werden. Dies betrifft zum einen den privaten Bereich, wenn beispielsweise mittels individualisierter Verhaltensprofile Vorhersagen über das Konsumverhalten getroffen werden, um dieses im Interesse der Anbieter zu beeinflussen.[21] Zum anderen ist auch der Bereich der politischen Willensbildung betroffen, etwa im Bereich von Mikrotargeting im Wahlkampf.[22] Diese und andere Beispiele machen deutlich, dass sich die Interessen von Verbraucherinnen und Verbrauchern auf keinen Fall darin erschöpfen, nur Adressaten von mit guten Daten trainierten KI-Systemen zu werden. Vielmehr kommt hier zum Tragen, dass eines der Ziele des Datenschutzrechts die Verhinderung zu starker Machtungleichgewichte ist; informationelle Selbstbestimmung ist gefährdet, wenn die betroffenen Personen nicht mehr wissen „wer was wann und bei welcher Gelegenheit über sie weiß",[23] und die Wissenden – also einflussreiche staatliche oder private Akteure – mittels KI-Algorithmen ihre Informationsmacht noch weiter ausbauen.

Unabhängig von der Frage, ob es sich um gute oder schlechte Trainingsdaten handelt, kann aus datenschutzrechtlicher Sicht auch das Training selbst ein Problem darstellen. Es geht dann nicht um Fragen des späteren Einsatzes eines mehr oder weniger gut trainierten KI-Systems in der Praxis, sondern um die vorgelagerte Verwendung personenbezogener Daten, einschließlich derjenigen solcher betroffenen Personen, die mit dem späteren KI-System gar nicht in Kontakt kommen. Die Risikolagen liegen dementsprechend auf einer anderen Ebene: Die ursprünglich zu anderen Zwecken erhobenen Daten werden nunmehr für das Training verarbeitet, länger aufbewahrt, an Forschungs- und Entwicklungspartner übermittelt oder sind schlimmstenfalls beim späteren Einsatz des KI-Systems erkennbar.[24] Diese Vorgänge können die Missbrauchsgefahren erheblich ansteigen lassen.

III. Betroffenenrechte in der Datenschutz-Grundverordnung

Datenschutzrecht gilt nur für personenbezogene Daten (Art. 2 Abs. 1 DSGVO, § 1 Abs. 1 S. 1 BDSG und entsprechende Normen der Landesdatenschutzgesetze). Daraus ergeben sich zwei Anwendungsfälle für die datenschutzrechtliche Perspektive auf KI-Trainingsdaten: zum einen das Training mit personenbezogenen Daten, zum anderen der Einsatz (wie auch immer) trainierter KI-Systeme in spä-

---

[21] S. etwa *Gausling*, ZD 2019, 335; zur Dynamisierung von Preisen etwa *Bernhardt*, NZKart 2019, 314; zu Transparenzanforderungen aus Verbrauchersicht *Gerpott/Mikolas*, InTeR 2021, 122.

[22] Dazu aus rechtlicher Sicht *Richter*, in: FS für Alexander Roßnagel, 2020, 303 ff.; *Radtke*, K&R 2020, 479.

[23] BVerfGE 65, 1 (43) – Volkszählung.

[24] Zu diesem Problem eines Personenbezugs von KI-Modellen s. *Kaulartz*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.9 Rn. 2 ff.

teren Anwendungen unter Einsatz personenbezogener Daten. In beiden Fällen können die Betroffenenrechte nach der Datenschutz-Grundverordnung eine Rolle spielen.

1. Die datenschutzrechtlichen Betroffenenrechte

Das Datenschutzrecht kennt schon seit vielen Jahren ein relativ festes Set von Betroffenenrechten. Bereits das erste Bundesdatenschutzgesetz aus dem Jahre 1977 umfasste Rechte auf Auskunft, Berichtigung, Sperrung und Löschung (verankert in § 4 i. V. m. §§ 13, 14, 26, 27, 32, 33 BDSG 1977). Auf europäischer Ebene enthielt die Datenschutz-Richtlinie 95/46/EG ebenfalls derartige Rechte (Art. 12) sowie Informationspflichten (Art. 10 und Art. 11), ein Widerspruchsrecht (Art. 14) und Rechte bei automatisierten Einzelentscheidungen (Art. 15).

Die Datenschutz-Grundverordnung hat diese Rechte im Grundsatz übernommen, die entsprechenden Bestimmungen jedoch stark ausgebaut, erweitert und präzisiert. Dies hat zu einer Fülle von neuen Streitfragen geführt, die an dieser Stelle nicht diskutiert werden können.[25] Im Überblick: Art. 12 DSGVO regelt nunmehr vor die Klammer gezogen Anforderungen an transparente Information, Kommunikation und Modalitäten für die Ausübung der Betroffenenrechte. Sodann folgen Informationspflichten für den Fall der Erhebung personenbezogener Daten bei der betroffenen Person (Art. 13 DSGVO) und bei anderen Stellen oder Gelegenheiten (Art. 14 DSGVO). Mit diesen proaktiven Informationspflichten korrespondiert das ebenfalls transparenzorientierte Auskunftsrecht in Art. 15 DSGVO.

Weiterhin enthalten sind in neu gefasster Form die Rechte auf Berichtigung (Art. 16 DSGVO), auf Löschung (Art. 17 DSGVO; nunmehr synonym – und gegenüber dem Norminhalt deutlich überschießend – „Recht auf Vergessenwerden" genannt), auf Einschränkung der Verarbeitung (Art. 18 DSGVO, vormals Sperrung), auf Widerspruch (Art. 21 DSGVO) sowie im Bereich der automatisierten Entscheidungen im Einzelfall einschließlich Profiling (Art. 22 DSGVO). Eine echte, das Datenschutzrecht in Richtung Wettbewerbs-, Kartell- und Verbraucherschutzrecht transzendierende Neuerung ist das Recht auf Datenübertragbarkeit (Art. 20 DSGVO).[26] Art. 23 DSGVO gestattet Beschränkungen der Betroffenen-

---

[25] S. nur zur Frage der Reichweite des Auskunftsrechts in Art. 15 DSGVO: *Kremer*, CR 2018, 560; *Zikesch/Sörup*, ZD 2019, 239; *Brink/Joos*, ZD 2019, 483; *Wybitul/Brams*, NZA 2019, 672; *Schulte/Welge*, NZA 2019, 1110; *Koreng*, NJW 2021, 2692; *Krämer/Burghoff*, ZD 2022, 428; *Gaul/Pitzer*, DB 2022, 1321; *Lembke/Fischels*, NZA 2022, 513. Die Norm beschäftigt inzwischen umfangreich die Justiz; s. nur aus der höchstrichterlichen Rechtsprechung BGH, NJW 2021, 2726; ZD 2022, 326; BVerwG, NVwZ 2021, 80; BAG, NJW 2021, 2379; BFHE 274, 496; DStRK 2020, 54; der BGH hat Fragen zum kostenfreien Auskunftsanspruch eines Patienten gegen seinen Arzt dem EuGH vorgelegt, s. BGH, GesR 2022, 360.

[26] Zu den Neuerungen der Betroffenenrechte schon *Hornung*, in: Hill/Schliesky (Hrsg.), Die Neubestimmung der Privatheit. E-Volution des Rechts- und Verwaltungssystems IV, 2014, 139 ff.; s. ferner *Reich*, VuR 2018, 293.

rechte durch Rechtsvorschriften der Union oder der Mitgliedstaaten, wenn die Beschränkungen den Wesensgehalt der Grundrechte und Grundfreiheiten achten sowie in einer demokratischen Gesellschaft notwendig und verhältnismäßig zur Sicherung enumerativ aufgelisteter gegenläufiger Interessen sind.

Im Grundsatz greifen alle diese Betroffenenrechte in beiden oben genannten KI-Szenarien, also sowohl beim Training von KI-Systemen mit personenbezogenen Daten als auch beim Einsatz von KI-Systemen in personenbezogenen Anwendungen.[27] Allerdings kommen nicht in jedem Fall die Besonderheiten von KI zu tragen. Die folgende Darstellung konzentriert sich deshalb auf einige zentrale Fragestellungen.

2. Ansprüche gegen die Verwendung zu Trainingszwecken

Im Trainingsszenario wäre aus verbraucherschutzrechtlicher Perspektive ein Anspruch gegen die Verwendung der „eigenen" personenbezogenen Daten zu Testzwecken das stärkste Betroffenenrecht. Da das Datenschutzrecht keinen expliziten Unterlassungsanspruch kennt, kommen derartige Ansprüche typischerweise im Gewande des Löschungsanspruchs daher.[28] Personenbezogene Daten sind insbesondere zu löschen, wenn sie unrechtmäßig verarbeitet werden (Art. 17 Abs. 1 lit. d DSGVO). Dies verweist auf das datenschutzrechtliche Verbotsprinzip,[29] rechtlich als Grundsatz der Rechtmäßigkeit in Art. 5 Abs. 1 lit. a und Art. 6 DSGVO verankert. Danach ist für jede Verarbeitung personenbezogener Daten eine sie legitimierende Rechtsgrundlage erforderlich.

a) Zulässigkeitstatbestände

Mangels eines expliziten Erlaubnistatbestands für die Verarbeitung zum Training von KI-Algorithmen richtet sich die datenschutzrechtliche Zulässigkeit nach den allgemeinen Regeln des Art. 6 DSGVO;[30] dies gilt jedenfalls vorbehaltlich einer

---

[27] S. näher *Niemann/Kevekordes*, CR 2020, 179, 181 ff.; *Krügel/Pfeiffenbring*, in: Ebers/Heinze/Krügel/Steinrötter (Fn. 13), § 11 Rn. 29 ff.

[28] Dies gilt auch für den EuGH, etwa in den Entscheidungen zum Recht auf Vergessenwerden (EuGH, Urt. v. 13.5.2014, Rs. C-131/12, ECLI:EU:C:2014:317 – Google Spain; Urt. v. 24.9.2019, Rs. C-136/17, ECLI:EU:C:2019:773 – CG u. a.; Urt. v. 24.9.2019, Rs. C-507/17, ECLI:EU:C:2019:772 – Google; dazu Hornung, in: FS für Alexander Roßnagel, 2020, 379 ff.). Die Frage, ob es – beispielsweise als Minus zum Löschungsanspruch nach Art. 17 DSGVO und insoweit in diesem verankert – einen allgemeinen datenschutzrechtlichen Unterlassungsanspruch gibt, kann hier nicht vertieft werden, s. dazu LG Frankfurt, ZD 2019, 410; LG Wiesbaden, MMR 2022, 313; *Leibold/Laoutoumai*, ZD-Aktuell 2021, 05583; *Worms*, in: BeckOK Datenschutzrecht, Art. 17 Rn. 77a f.

[29] S. z. B. *Karg*, DuD 2013, 75 ff.; kritisch zur Terminologie *Roßnagel*, NJW 2019, 1.

[30] S. dazu *Niemann/Kevekordes*, CR 2020, 17, 22 ff.; allgemeiner zur Zulässigkeit der Datenverarbeitung zu Testzwecken *Heinemeyer*, CR 2019, 147; allgemeiner für KI-Systeme *Conrad*, InTeR 2021, 147 (v. a. zur Einwilligung). Für wissenschaftliche Forschung gelten Sonderregeln, die hier ausgeklammert werden, s. z. B. *Meszaros/Ho*, Computer Law & Security Review 41 (2021), 105532.

Sperrwirkung von Art. 54 KOM-E (s. u. III 2. d). Von den dort genannten Alternativen dürften etliche nur selten einschlägig werden. Denn die Verarbeitung personenbezogener Daten zum Training von KI-Systemen wird nur in Ausnahmefällen zur Erfüllung einer rechtlichen Verpflichtung (lit. c), zum Schutz lebenswichtiger Interessen (lit. d) oder zur Wahrnehmung einer Aufgabe im öffentlichen Interesse oder in Ausübung öffentlicher Gewalt (lit. e) erforderlich sein. Die Erforderlichkeit zur Erfüllung eines Vertrags (Art. 6 Abs. 1 UAbs. 1 lit. b DSGVO) kommt in Betracht, wenn das Training selbst Vertragsgegenstand ist oder es um eine personalisierte Dienstleistung geht, die nur auf der Basis eines Trainings mit den personenbezogenen Daten der betroffenen Person angeboten werden kann. Allerdings ist die Rechtsgrundlage sodann auf das Training mit den personenbezogenen Daten des Vertragspartners beschränkt; die Verarbeitung von Daten Dritter kann nicht auf Art. 6 Abs. 1 UAbs. 1 lit. b DSGVO gestützt werden.

In den meisten Fällen wird es um die Frage einer Einwilligung (Art. 6 Abs. 1 UAbs. 1 lit. a i. V. m. Art. 7 und Art. 4 Nr. 11 DSGVO), das Problem gleichgewichtiger oder überwiegender berechtigter Interessen eines privaten Verantwortlichen oder eines Dritten (Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO) bzw. Verhältnismäßigkeitsüberlegungen im Rahmen von Verarbeitungsgeneralklauseln für den öffentlichen Bereich gehen (Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 2 und 3 DSGVO i. V. m. § 3 BDSG bzw. entsprechenden Normen der LDSGe). Der Weg über Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO versagt freilich, soweit es um besondere Kategorien personenbezogener Daten nach Art. 9 Abs. 1 DSGVO geht.[31] Denn dann greift das dort geregelte zusätzliche Verarbeitungsverbot, das nur in den Fällen des Art. 9 Abs. 2 DSGVO überwunden werden kann. Dessen Ausnahmetatbestände enthalten keine allgemeine Interessenabwägung wie in Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO. Die Einwilligung kann hingegen weiterhin verwendet werden, muss nach Art. 9 Abs. 2 lit. a DSGVO aber ausdrücklich erteilt werden. Die übrigen Ausnahmetatbestände des Art. 9 Abs. 2 DSGVO können in einigen Fällen für die Verarbeitung als KI-Trainingsdaten einschlägig sein (zum Beispiel lit. e, wenn die betroffene Person die Daten offensichtlich selbst öffentlich gemacht hat), dies ist aber im Einzelfall zu beurteilen.

b) Zweckänderung nach der DSGVO

Im Rahmen der erforderlichen Einzelfallprüfung sind verschiedene Szenarien zu unterscheiden. Gerade mit einer Einwilligung nach Art. 7 i. V. m. Art. 4 Nr. 11 DSGVO kann der Verantwortliche Daten zum originären Zweck des Trainings

[31] Dies umfasst personenbezogener Daten, aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie genetische Daten, biometrische Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung. Zum Problem der Anwendung auf Machine Learning s. *Niemann/Kevekordes*, CR 2020, 179.

von KI-Algorithmen erheben und weiterverarbeiten. Unter bestimmten Voraussetzungen ist aber auch die Zweckänderung für Daten möglich, die zunächst zu einem anderen Zweck erhoben wurden. Schließlich kann es auch darum gehen, Daten an andere Verantwortliche weiterzugeben, um beispielsweise das Wissen über selten auftretende Anomalien (etwa Unfallursachen im Straßenverkehr) mit anderen zu teilen und diesen ebenfalls die Möglichkeit zu geben, ihre Algorithmen entsprechend zu trainieren.

In der Praxis wird es oftmals um den Fall der Zweckänderung gehen, wenn beispielsweise bestehende Kundendatenbanken, Sammlungen mit Behandlungsinformationen von Patienten oder Verwaltungsakten bereits über einen längeren Zeitraum angelegt wurden und ihr Einsatz zum Training neuer Algorithmen nunmehr als vielversprechend erscheint.[32] Der in Art. 5 Abs. 1 lit. b DSGVO geregelte Grundsatz der Zweckbindung enthält eine Privilegierung der Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke; diese Verarbeitung gilt als nicht unvereinbar mit dem ursprünglichen Zweck, wenn Sicherungsmittel nach Art. 89 Abs. 1 DSGVO ergriffen werden.[33] Jenseits dieser Fälle ist jedoch eine Prüfung erforderlich, ob der neue Verarbeitungszweck mit dem ursprünglichen „vereinbar" ist. Hierfür gibt Art. 6 Abs. 4 DSGVO Kriterien vor.

Auch diese Vorschrift wirft etliche Probleme auf.[34] Für die hier diskutierte Konstellation ist insbesondere wichtig, dass die risikobezogenen Abwägungskriterien nur teilweise so verallgemeinert werden können, dass sie pauschale Argumente für die Zweckänderung umfangreicher Datenbestände liefern. Dies trifft beispielsweise auf die geeigneten Garantien wie Verschlüsselung oder Pseudonymisierung zu (Art. 6 Abs. 4 lit. d DSGVO), weil diese durch den Verantwortlichen einheitlich angewendet und dementsprechend einheitlich in die Abwägung eingebracht werden können. Es gilt aber bereits nur noch mit Einschränkungen – d. h. bei ähnlich strukturierten Daten – für die Kriterien der Verbindung zwischen altem und neuem Zweck (lit. a), des Erhebungszusammenhangs (lit. b) und die Frage, ob besondere Kategorien personenbezogener Daten nach Art. 9 DSGVO oder über strafrechtliche Verurteilungen und Straftaten nach Art. 10 DSGVO verarbeitet werden (lit. c). Hier können sich bereits deutliche Unterschiede zwischen ein-

---

[32] *Niemann/Kevekordes*, CR 2020, 17, 24; *Valkanova*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.1 Rn. 3 ff.

[33] Zu dieser Privilegierung s. *Werkmeister/Schwaab*, CR 2019, 85; *Weichert*, ZD 2020, 18; am Beispiel von Patientendaten *Spitz/Jungkunz/Schickhardt/Cornelius*, MedR 2021, 499; allgemeiner zum Datenschutz in der Forschung *Roßnagel*, ZD 2019, 157.

[34] Dies betrifft schon die grundsätzliche Frage, ob im Falle der Einschlägigkeit von Art. 6 Abs. 4 DSGVO die Weiterverarbeitung sich auf die ursprüngliche Rechtsgrundlage stützt (dafür z. B. *Roßnagel*, in Simitis/Hornung/Spiecker gen. Döhmann (Hrsg.), Datenschutzrecht, 2019, Art. 6 Abs. 4 Rn. 12 m. w. N.; *Kühling/Martini*, EuZW 2016, 448, 451; *Hornung/Hofmann*, ZD-Beilage 4/2017, 8) oder zusätzlich eine weitere Rechtsgrundlage nach Art. 6 Abs. 1 DSGVO erforderlich ist (so *Schantz*, NJW 2016, 1841, 1844; *Albrecht*, CR 2016, 88, 92).

zelnen betroffenen Personen ergeben. Die möglichen Folgen der beabsichtigten Weiterverarbeitung für die betroffene Person (Art. 6 Abs. 4 lit. d DSGVO) werden sogar oftmals eine Betrachtung ihrer konkreten Umstände erfordern. Im Ergebnis wird es deshalb schwierig sein, die einheitliche Zweckänderung für große Datenbestände nach diesen Kriterien übergreifend zu legitimieren. Zu diesem Problem der Einzelfallabhängigkeit tritt hinzu, dass die Sensibilität der Daten gerade im Anwendungsbereich von Art. 9 DSGVO (beispielsweise bei Gesundheitsdaten) die Zweckänderung deutlich einschränkt.

c) Zweckänderung im KOM-E

Bedeutsam ist deshalb, dass Art. 6 Abs. 4 DSGVO neben der Zweckänderung aufgrund des erläuterten Vereinbarkeitstests auch die Möglichkeit vorsieht, dass die Zweckänderung durch eine Einwilligung der betroffenen Person oder eine Rechtsvorschrift der Union oder der Mitgliedstaaten legitimiert wird. Für entsprechende Normen gelten dieselben Anforderungen wie für eine Einschränkung der Betroffenenrechte nach Art. 23 DSGVO (s. o.). Solche Rechtsvorschriften schlägt die Kommission in Art. 10 Abs. 5 KOM-E sowie in Art. 54 Abs. 1 KOM-E vor.[35]

aa) Vermeidung von Verzerrungen

Nach Art. 10 Abs. 5 KOM-E dürfen Anbieter von Hochrisiko-KI-Systemen Daten nach Art. 9 Abs. 1 DSGVO[36] verarbeiten, soweit dies für die Beobachtung, Erkennung und Korrektur von „Verzerrungen" im Zusammenhang mit Hochrisiko-KI-Systemen unbedingt erforderlich ist. Dies ist ein hoher Maßstab, der über die übliche Erforderlichkeitsprüfung im Datenschutzrecht hinausreicht. Die Anbieter müssen außerdem angemessene Vorkehrungen für den Schutz der Grundrechte und Grundfreiheiten natürlicher Personen treffen, wozu auch technische Beschränkungen einer Weiterverwendung und modernste Sicherheits- und Datenschutzmaßnahmen wie Pseudonymisierung oder Verschlüsselung gehören,

---

[35] Im Folgenden wird die Frage ausgeklammert, ob die vorgeschlagenen Regelungen tatsächlich eines der Ziele in Art. 23 Abs. 1 DSGVO verfolgen. Für Art. 10 Abs. 5 KOM-E kommt Art. 23 Abs. 1 lit. i DSGVO in Betracht, da Verzerrungen die Rechte und Freiheiten natürlicher Personen beeinträchtigen können. Letztlich spielt die Frage aber keine Rolle, weil die Regelungen des KOM-E spezieller sind und (entgegen der Formulierung in Art. 6 Abs. 4 DSGVO) nicht den Vorgaben der DSGVO folgen müssen. Diese entsprechen zwar hinsichtlich der Verhältnismäßigkeitsanforderungen im Wesentlichen den Vorgaben in Art. 52 Abs. 1 GRCh. Dieser kennt aber keine enumerative Liste zulässiger Gemeinwohlziele.

[36] Zusätzlich werden Daten nach Art. 10 RL (EU) 2016/680 und Art. 10 Abs. 1 VO (EU) 2018/1725 genannt. Diese regeln allerdings dieselben Daten und unterscheiden sich nur hinsichtlich des persönlichen Anwendungsbereichs (Strafverfolgungsbehörden bzw. Organe, Einrichtungen und sonstige Stellen der Union; beides ist nach Art. 2 Abs. 2 lit. d und Abs. 3 DSGVO von dieser ausgenommen). Da der persönliche Anwendungsbereich allerdings in Art. 10 Abs. 5 KOM-E ohnehin anders bestimmt wird, erschließt sich der Mehrfachverweis nicht.

wenn der verfolgte Zweck durch eine Anonymisierung erheblich beeinträchtigt würde. Mit der Verpflichtung auf „modernste" Maßnahmen wird ein sehr hoher Standard vorgegeben, der über den Stand der Technik hinausreichen und dem aus dem deutschen Recht bekannten Stand von Wissenschaft und Technik entsprechen wird.[37]

Art. 10 Abs. 5 KOM-E zielt erkennbar auf die Vermeidung von Ungleichheitseffekten und Diskriminierungen sowie darüber hinaus auf die allgemeine Qualität von Entscheidungs(unterstützung)systemen mithilfe von KI. Denn um herauszufinden, ob ein KI-System Menschen beispielsweise nach ihrer rassischen und ethnischen Herkunft, ihren genetischen oder biometrischen Daten oder ihrer Gesundheit ungleich behandelt, wird man – vorbehaltlich der Verfügbarkeit synthetischer Trainingsdaten – typischerweise nicht umhinkommen, personenbezogene Daten zu genau diesen Merkmalen zu verarbeiten.[38] Der Vorschlag entbindet als gesetzliche Verarbeitungsbefugnis die Anbieter davon, von jeder betroffenen Person eine ausdrückliche Einwilligung einzuholen. Um diese Einschränkung der Rechtsposition von Verbraucherinnen und Verbrauchern auch grundrechtlich rechtfertigen zu können, hat die Kommission hohe technische und organisatorische Kompensationsmaßnahmen vorgesehen. Insofern ist die Vorschrift ein sinnvoller Regelungsansatz, wirft in der vorgeschlagenen Form allerdings dennoch mehrere Fragen auf.

Erstens ist nicht eindeutig, was mit „Verzerrungen" gemeint ist. Nach Art. 10 Abs. 2 lit. f KOM-E müssen Daten-Governance- und Datenverwaltungsverfahren eine Untersuchung „im Hinblick auf mögliche Verzerrungen (Bias)" umfassen. Es geht also um Schieflagen, Bevorzugungen und Benachteiligungen sowie diskriminierende Effekte – der Bezug zwischen Verzerrungen und Diskriminierungen kommt in EG 33 S. 1 und besonders deutlich in EG 44 S. 6 KOM-E zum Ausdruck.[39] Damit ist die Zielrichtung der Regelung umschrieben. Was genau (relevante) Verzerrungen ausmacht und umfasst, verbleibt dennoch deutlich im Unklaren.

Zweitens erstreckt sich Art. 10 Abs. 5 KOM-E nur auf Daten nach Art. 9 Abs. 1 DSGVO. Es bleibt also offen, inwieweit die Anbieter zur Vermeidung von Verzerrungen auch „einfache" personenbezogene Daten verarbeiten dürfen. Für einen Umkehrschluss, also ein Verarbeitungsverbot, gibt es keinen Anhaltspunkt im KOM-E; ein solches Verbot der Verarbeitung einfacher Daten wäre angesichts der Verarbeitungsbefugnis für sensible Daten auch nicht zu rechtfertigen. In diese Richtung deutet auch EG 44 S. 6 KOM-E, demzufolge die Anbieter angesichts des

---

[37] S. dazu allgemein BVerfGE 49, 89 (136); BVerfG, NJW 1980, 759, 761 f.; BVerwGE 92, 185 (196); BVerwGE 106, 115; *Seibel*, NJW 2013, 3000, 3003; aus datenschutzrechtlicher Perspektive *Bartels*/*Backer*, DuD 2018, 214, 215.

[38] Dazu *Žliobaitė*/*Custers*, Artificial Intelligence and Law 24 (2016), 183.

[39] S. a. das Gutachten der *Datenethikkommission*, 2019, 231.

erheblichen öffentlichen Interesses „auch" besondere Kategorien personenbezogener Daten verarbeiten dürfen, um Verzerrungen in Hochrisiko-KI-Systemen zu beobachten, zu erkennen und zu korrigieren. Die Kommission geht also offenbar davon aus, dass andere Daten auf Basis anderer Rechtsgrundlagen ebenfalls verarbeitet werden können. Gegen den Willen der betroffenen Personen kommt dies insbesondere auf Basis einer Abwägung nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO bzw. im Rahmen von Generalklauseln wie Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 2 und 3 DSGVO i. V. m. § 3 BDSG in Betracht (s. o.). Im Rahmen der entsprechenden Abwägungen sind zwar wie bei Art. 10 Abs. 5 KOM-E technische und organisatorische Maßnahmen des Verantwortlichen zu berücksichtigen.[40] Es wäre aber vorzugswürdig, wenn der Gesetzgeber die Verarbeitungsbefugnis und spezifische, KI-bezogene Abwägungskriterien auch für normale personenbezogene Daten in der Verordnung regeln würde.

Eine ähnliche Frage ergibt sich drittens aus der Beschränkung des Vorschlags auf Hochrisiko-KI-Systeme. Verzerrungseffekte können auch bei KI-Systemen eintreten, die nicht in diese Kategorie fallen. Da es sich bei der Risikokategorisierung um eine typisierende Einordnung handelt, kann die Verzerrung in atypischen Einzelfällen für Verbraucherinnen und Verbraucher durchaus gravierende Folgen haben, obwohl es sich nicht um ein Hochrisiko-KI-System handelt. Gegen eine Ausweitung der Verarbeitungsbefugnis in Art. 10 Abs. 5 KOM-E auf alle KI-Systeme lässt sich (sowohl rechtspolitisch als auch hinsichtlich einer Analogie) allerdings anführen, dass der Bedarf nach einer Verarbeitung von Daten nach Art. 9 Abs. 1 DSGVO bei Hochrisiko-KI-Systemen deutlich größer ist, da diese eben mit typisiert erhöhten Risiken für Verbraucherinnen und Verbraucher verbunden sind. Insofern handelt es sich weniger um eine Abwägung zwischen den Interessen der Anbieter bzw. Nutzer des KI-Systems im Verhältnis zu den betroffenen Personen, sondern um eine Abwägung zwischen den Interessen der späteren Adressaten des Hochrisiko-KI-Systems einerseits und der betroffenen Personen andererseits, deren Daten zur Vermeidung von Verzerrungen verarbeitet werden sollen. Angesichts dieser Gemengelage sollte der europäische Gesetzgeber die Frage aber explizit entscheiden. Dies gilt insbesondere, weil eine Analogie zu Art. 10 Abs. 5 KOM-E für nicht erfasste, „normale" KI-Systeme aufgrund des risikoklassenorientierten Regelungsansatzes und der detaillierten Vorschläge der Kommission zwar zweifelhaft, methodisch aber immerhin möglich wäre.

bb) KI-Reallabore

Art. 54 Abs. 1 KOM-E enthält eine detaillierte Befugnis zur Zweckänderung von Daten, die rechtmäßig für andere Zwecke erhoben wurden; sie ist dementspre-

[40] S. allgemein zur Abwägung nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO z. B. *Robrahn/Bremert*, ZD 2018, 291; *Herfurth*, ZD 2018, 514.

chend ausweislich EG 72 S. 4 KOM-E ausdrücklich als Regelung im Sinne von Art. 6 Abs. 4 DSGVO zu sehen. Diese dürfen in einem „KI-Reallabor" verarbeitet werden, um innovative KI-Systeme zu entwickeln oder zu erproben, wenn diese ein erhebliches öffentliches Interesse im Bereich der Straftatbekämpfung, der öffentlichen Sicherheit und öffentlichen Gesundheit oder des Umweltschutzes verfolgen. Anders als Art. 10 Abs. 5 KOM-E ist die Regelung nicht auf Daten nach Art. 9 Abs. 1 DSGVO beschränkt, sondern gestattet umfassend die Verarbeitung der personenbezogenen Daten von Verbraucherinnen und Verbrauchern als datenschutzrechtlich betroffene Personen.[41] Art. 54 Abs. 1 KOM-E enthält sodann eine Vielzahl insbesondere organisatorischer Anforderungen an die Datenverarbeitung sowie eine Löschpflicht, um die Grundrechte der betroffenen Personen zu schützen.

Die Regelung stellt im Ergebnis erkennbar hohe Anforderungen an die Zweckänderung. Welche Relevanz sie erlangen wird, hängt maßgeblich davon ab, wie die konkrete Ausgestaltung der KI-Reallabore ausfallen wird. Nach Art. 53 Abs. 1 KOM-E werden sie von den zuständigen Behörden eines oder mehrerer Mitgliedstaaten oder vom europäischen Datenschutzbeauftragten eingerichtet. Sie bieten eine kontrollierte Umgebung, um die Entwicklung, Erprobung und Validierung innovativer KI-Systeme für einen begrenzten Zeitraum vor ihrem Inverkehrbringen oder ihrer Inbetriebnahme nach einem „spezifischen Plan" zu erleichtern.[42] Die Modalitäten und Bedingungen für den Betrieb der KI-Reallabore werden allerdings erst nach Art. 53 Abs. 6 KOM-E in Durchführungsrechtsakten der Kommission festgelegt. Ob Entwickler und Anbieter von KI-Systemen von der Option einer Entwicklung unter „direkter Aufsicht und Anleitung der zuständigen Behörden" (Art. 53 Abs. 1 S. 2 KOM-E) Gebrauch machen werden, dürfte maßgeblich von dieser konkreten Ausgestaltung abhängen.

d) Exkurs: Zweckänderung nach § 8 Abs. 1 und 2 ITEG SH

Eine deutlich weitergehende Befugnis zur Zweckänderung als in Art. 10 Abs. 5 und Art. 54 Abs. 1 KOM-E enthält § 8 ITEG SH. § 8 Abs. 1 S. 1 ITEG SH legitimiert die Verarbeitung von (personenbezogenen und nicht personenbezogenen) Daten durch Träger der öffentlichen Verwaltung zum Zweck der Entwicklung und des Trainings von „datengetriebenen Informationstechnologien" (dies sind der Sache nach KI-Systeme).[43] § 8 Abs. 1 S. 2 ITEG SH beschränkt die Verarbeitung perso-

---

[41] Kritisch *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188, 1192.

[42] S. zu diesen „regulatory sandboxes" *Spindler*, CR 2021, 361, 371.

[43] § 3 Abs. 1 Nr. 1 ITEG SH definiert datengetriebene Informationstechnologien als Basisdienste, Fachverfahren oder Fachanwendungen, die zur effizienten Lösung einer speziellen Aufgabe oder einer komplexen Fragestellung auf Grundlage eines Datensatzes mit Hilfe spezieller Systeme, wie künstlicher neuronaler Netze und maschineller Lernverfahren, eingesetzt werden und ohne aktiven Eingriff Parameter der Entscheidungsfindung weiterentwickeln. Das Gesetz verwendet also nicht den Begriff der KI, regelt diese aber der Sache nach. Dies ist auch Absicht des Gesetzgebers, s. ausdrücklich LT-Drs. SH 19/3267, 15 f.

nenbezogener Daten zu Zwecken des Trainings (nicht der Entwicklung) sodann auf solche Fälle, in denen ein effektives Training nur mit unverhältnismäßigem Aufwand auf andere Weise erfolgen kann. Dies dürfte der Sache nach dem verschärften Maßstab der unbedingten Erforderlichkeit in Art. 10 Abs. 5 KOM-E entsprechen.

§ 8 Abs. 2 S. 1 ITEG SH beschränkt sodann den Kreis derjenigen Daten, die einer Zweckänderung unterzogen werden dürften. Wenn personenbezogene Daten zu Trainingszwecken verarbeitet werden sollen oder nicht auszuschließen ist, dass personenbezogene Daten betroffen sein könnten, so dürfen nur solche Daten verarbeitet werden, die im Zusammenhang mit der zu trainierenden Aufgabenwahrnehmung erhoben und gespeichert wurden. Ein Datenaustausch über verschiedene Bereiche der öffentlichen Verwaltung hinweg ist damit ausgeschlossen oder zumindest auf solche Bereiche beschränkt, bei denen dieselbe Aufgabe erfüllt wird. § 8 Abs. 2 S. 2 ITEG SH enthält die Vorgabe, die Daten vor einer Verarbeitung zu Trainingszwecken zu pseudonymisieren, sofern der Zweck dadurch nicht verhindert wird. Dies ist in § 10 Abs. 5 KOM-E ebenfalls enthalten, dort werden aber weitergehende technische und organisatorische Maßnahmen angeordnet (s. o.).

Art. 10 Abs. 5 KOM-E und § 8 ITEG SH weisen zwei entscheidende Unterschiede auf. Zum einen ist Art. 10 Abs. 5 KOM-E – wie die meisten Vorgaben des Entwurfs – auf Hochrisiko-KI-Systeme beschränkt, während § 8 ITEG SH die Zweckänderung für alle KI-Systeme in der öffentlichen Verwaltung legitimiert. Zum anderen ist § 8 ITEG SH nicht auf den Zweck der Vermeidung von Verzerrungen beschränkt, sondern betrifft allgemein die Verarbeitung zu Zwecken der Entwicklung und des Trainings von KI-Systemen (die Verarbeitung beim späteren Einsatz richtet sich nach allgemeinen Regeln). Dies entspricht weitgehend der Zielrichtung von Art. 54 Abs. 1 KOM-E für die Entwicklung und Erprobung. Allerdings ist diese Norm viel enger als § 8 ITEG SH, weil erhebliche Anforderungen an die Verarbeitung durch KI-Reallabore festgeschrieben werden, die weit über die im Vergleich rudimentären Vorgaben in § 8 ITEG SH hinausgehen.

Sollte der Verordnungsentwurf in dieser Form beschlossen werden, wird deshalb die Frage zu beantworten sein, ob die unionsrechtliche Zweckänderungsbefugnis nationale Regelungen sperrt, die an sich auf Basis der Öffnungsklauseln der DSGVO zulässig wären. Hierfür könnte man anführen, dass andernfalls die hohen Vorgaben von Art. 54 Abs. 1 KOM-E sehr leicht unterlaufen werden könnten. Andererseits sind die dort geregelten KI-Reallabore ein sehr spezieller Anwendungsfall. Da sie bisher noch nicht einmal konzeptionell wirklich existieren (Modalitäten und Bedingungen für den Betrieb werden nach Art. 53 Abs. 6 KOM-E in Durchführungsrechtsakten festgelegt, s. o.), würde die Annahme einer Sperrwirkung die allermeisten Anwendungsfälle einer Zweckänderung personenbezogener Daten zur Entwicklung und Erprobung von KI-Systemen verbieten. Für eine solche Wirkung des Entwurfs bietet der KOM-E weder in EG 72 noch

an sonstiger Stelle einen Anhaltspunkt; ohne einen solchen sollte eine so weitgehende Rechtsfolge nicht angenommen werden.

3. Transparenzpflichten

Neben der Zulässigkeit der Datenverarbeitung und einem etwaigen Abwehranspruch gegen diese spielt die Transparenz eine zentrale Rolle sowohl im Verbraucherschutz- als auch im Datenschutzrecht. Das Bundesverfassungsgericht hat dies bereits im Volkszählungsurteil mit der eingängigen Formulierung hervorgehoben, mit dem Recht auf informationelle Selbstbestimmung, „wären eine Gesellschaftsordnung und eine diese ermöglichende Rechtsordnung nicht vereinbar, in der Bürger nicht mehr wissen können, wer was wann und bei welcher Gelegenheit über sie weiß".[44]

a) Transparenzpflichten bei der Zweckänderung zu Testzwecken

Sekundärrechtlich ist der Transparenzgrundsatz in Art. 5 Abs. 1 lit. a DSGVO verankert und wird maßgeblich durch das Auskunftsrecht (Art. 15 DSGVO) und die Informationspflichten der Art. 13, 14 DSGVO konkretisiert.[45] Beide Vorschriften verpflichten nicht nur bei der Datenerhebung dazu, die betroffene Personen zu informieren (Art. 13 Abs. 1, Art. 14 Abs. 1 DSGVO). Darüber hinaus muss nach Art. 13 Abs. 3 und Art. 14 Abs. 4 DSGVO ein Verantwortlicher, der beabsichtigt, die personenbezogenen Daten für einen anderen Zweck weiterzuverarbeiten als den, für den sie erlangt wurden, der betroffenen Person vor der Weiterverarbeitung Informationen über diesen anderen Zweck und alle anderen maßgeblichen Informationen gemäß dem jeweiligen Abs. 2 der Vorschrift zur Verfügung stellen.[46] Dies gilt im Ausgangspunkt für alle oben erwähnten Zweckänderungsbefugnisse im geltenden Recht (vor allem Art. 6 Abs. 4 DSGVO und § 8 ITEG SH) und würde auch für die Zweckänderungen nach dem KOM-E gelten.

Wurden die Daten nicht bei der betroffenen Person erhoben, so kann die Informationspflicht entfallen oder durch öffentliche Informationen zum Beispiel auf einer Webseite ersetzt werden, wenn die Erteilung der Informationen sich als unmöglich erweist oder einen unverhältnismäßigen Aufwand erfordern würde (Art. 14 Abs. 5 lit. b DSGVO).[47] Im Falle der Datenerhebung bei der betroffenen

---

[44] BVerfGE 65, 1 (43).

[45] S. zum Transparenzgrundsatz ausführlich *Manthey*, Das Datenschutzrechtliche Transparenzgebot, 2020; zur Umsetzung im Bereich von KI s. *Gausling*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.3; zum Problem der (fehlenden) Nachvollziehbarkeit von KI-Systemen noch näher unten IV 1.

[46] Zu dieser Transparenz bei Zweckänderung s. *Bäcker*, in: Kühling/Buchner (Hrsg.), DSGVO BDSG, 3. Auflage 2020, Art. 13 Rn. 69 ff.; Paal/Hennemann, in: Paal/Pauly (Hrsg.), DSGVO BDSG, 3. Auflage 2021, Art. 13 Rn. 33, jeweils m. w. N.

[47] Zur Auslegung und Reichweite dieser Ausnahme s. *Bäcker*, in: Kühling/Buchner (Fn. 46), Art. 14 Rn. 53 ff.; *Hennemann*, in: Paal/Pauly (Fn. 46), Art. 14 Rn. 40 ff., jeweils m. w. N.

Person existiert in Art. 13 DSGVO keine derartige Ausnahme. Einschränkungen sind allerdings auf der Basis von Art. 23 DSGVO zulässig (s. o.). Hiervon hat der deutsche Gesetzgeber v. a. in §§ 32, 33 BDSG Gebrauch gemacht.[48] Diese Normen dürften allerdings höchstens in sehr speziellen Einzelfällen der Zweckänderung von Daten zum Training von KI-Systemen greifen. Im Grundsatz werden deshalb die Rechte von Verbraucherinnen und Verbrauchern zumindest insoweit gewahrt, als ihnen eine solche Zweckänderung bereits vor der Weiterverarbeitung mitzuteilen ist.

b) Weitergabe der Trainingsdaten an andere betroffene Personen?

Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g DSGVO enthalten die Pflicht, zumindest in Fällen einer automatisierten Entscheidungsfindung einschließlich Profiling nach Art. 22 DSGVO „aussagekräftige Informationen über die involvierte Logik" sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person zur Verfügung zu stellen. Diese in ihrer Reichweite stark streitige Vorschrift[49] greift allerdings im Moment der Zweckänderung der Daten zum Zwecke des Trainings von KI-Systemen nicht, da dieses Training nicht in eine automatisierte Entscheidungsfindung für die betroffenen Personen mündet, deren Daten verarbeitet werden. Bei der späteren automatisierten Entscheidungsfindung unter Verwendung eines so trainierten KI-Systems ist die Regelung hingegen anwendbar.

Ob die Information über die involvierte Logik auch eine Information über die Trainingsdaten einschließt, ist offen. Die bisherige Rechtsprechung berücksichtigt zur Einschränkung des Auskunftsanspruchs insbesondere gegenläufige Betriebs- und Geschäftsgeheimnisse der Anbieter hinsichtlich der verwendeten Algorithmen.[50] Analog dazu wird es jedenfalls unzulässig sein, personenbezogene Trainingsdaten aus einer Phase der Entwicklung oder Implementierung des KI-Systems an eine andere natürliche Person herauszugeben, die nach den Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g DSGVO zu informieren ist, weil das fertige KI-System eine auf sie bezogene automatisierte Entscheidungsfindung vornimmt. Hierfür spricht auch, dass die Regelungen systematisch keine Zweckänderungs-

[48] S. zur Problematik dieser Einschränkungen und ihrer DSGVO-Konformität *Dix*, in Simitis/Hornung/Spiecker gen. Döhmann (Fn. 34), Art. 13 Rn. 23.

[49] Der BGH hat zu § 34 S. 1 Nr. 4 BDSG a. F. entschieden, dass der Auskunftsanspruch nicht die sogenannte Scoreformel, also die abstrakte Methode der Scorewertberechnung, umfasst, s. BGHZ 200, 38 (Rn. 27); s. a. Fn. 2. Es spricht viel dafür, dass die DSGVO über das Auskunftsrecht nach dem BDSG a. F. hinausgeht (a. A. wohl VG Wiesbaden, VuR 2022, 70), da die Formulierung zur involvierten Logik neu ist. In welchem Umfang Informationen über diese Logik bereitzustellen sind, ist aber unklar, s. näher *Kumkar/Roth-Isigkeit*, JZ 2020, 277, 281 ff.; *Sesing*, MMR 2021, 288.

[50] S. Fn. 49.

bzw. Übermittlungsbefugnisse der Verantwortlichen sind.[51] Demgegenüber ist es vorstellbar, dass Informationen allgemeiner Art über die verwendeten Trainingsdaten bereitzustellen sind, wenn sich hieraus die entscheidenden Informationen ergeben, um beispielsweise die Auswirkungen der Verarbeitung für die betroffene Person transparent zu machen.

c) Trainingsdaten als Teil der Dokumentation

Nach Art. 18 und Art. 16 lit. c KOM-E muss der Anbieter eines Hochrisiko-KI-Systems eine technische Dokumentation nach Art. 11 Abs. 1 KOM-E erstellen, bevor es in Verkehr gebracht oder in Betrieb genommen wird. Die Dokumentation enthält nach Art. 11 Abs. 1 UAbs. 2 S. 2 i. V. m. Anhang IV Nr. 2 lit. d und g KOM-E auch Angaben über Trainings-, Validierungs- und Testdaten. Aufzunehmen sind:

- Datenanforderungen in Form von Datenblättern, in denen die Trainingsmethoden und -techniken und die verwendeten Trainingsdatensätze beschrieben werden, mit Angaben zu Herkunft, Umfang und Hauptmerkmalen dieser Datensätze; Angaben zur Beschaffung und Auswahl der Daten; Kennzeichnungsverfahren (z. B. für überwachtes Lernen), Datenbereinigungsmethoden (z. B. Erkennung von Ausreißern);
- verwendete Validierungs- und Testverfahren, mit Angaben zu den verwendeten Validierungs- und Testdaten und deren Hauptmerkmalen;
- Parameter, die zur Messung der Genauigkeit, Robustheit, Cybersicherheit und der Erfüllung anderer einschlägiger Anforderungen nach Titel III Kapitel 2 sowie potenziell diskriminierender Auswirkungen verwendet werden;
- Testprotokolle und alle von den verantwortlichen Personen datierten und unterzeichneten Testberichte, auch in Bezug auf die in Buchstabe f genannten vorab bestimmten Änderungen.

Diese Angaben enthalten nicht die Trainings-, Validierungs- und Testdaten selbst,[52] sodass sich insoweit keine datenschutzrechtlichen Probleme ergeben; ähnliches dürfte für die Dokumentationspflicht in § 8 Abs. 4 ITEG SH gelten.[53] Demgegen-

---

[51] Systematisch wäre zumindest zu verlangen, dass zu den Tatbestandsvoraussetzungen der Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g DSGVO eine datenschutzrechtliche Befugnis zur (ggf. erneuten) Zweckänderung der personenbezogenen Trainingsdaten eingreift. Hierbei müssten die Grundrechte und Grundfreiheiten der betroffenen Personen berücksichtigt werden.

[52] Forderung nach einer Dokumentation dieser Daten in der Entschließung des Europaparlaments v. 6.10.2021, P9_TA(2021)0405, Rn. 19.

[53] Gemäß § 8 Abs. 4 S. 1 ITEG SH dokumentiert die öffentliche Stelle zwar „die für die Entwicklung und den Einsatz der datengetriebenen Informationstechnologien verwendeten Daten". Dies umfasst sprachlich auch die Daten selbst. S. 2 verlangt sodann aber, dass mindestens die Quelle der Daten, der Datenlieferant, der Erhebungskontext und der Erhebungszeitpunkt zu dokumentieren sind, nach S. 3 „sollen" (nur) „soweit möglich" auch die Mess- und Erhebungsmethode dokumentiert werden. Die Dokumentation der Daten selbst ist also jedenfalls nicht zwingend, entgegen dem Wortlaut von S. 1 mutmaßlich sogar nicht durch den Gesetzgeber beabsichtigt.

über sind Angaben über diese Daten sowie Informationen enthalten, die für die Kontrolle diskriminierender Auswirkungen auf Verbraucherinnen und Verbraucher relevant sein können. Die Dokumentation nach Art. 11 KOM-E ermöglicht somit auch eine auf die Trainings-, Validierungs- und Testdaten bezogene Kontrolle durch Dritte.

Allerdings erstreckt sich diese Kontrollmöglichkeit nicht auf die späteren Adressaten von KI-Systemen. Die Dokumentation dient zwar der Einhaltung der Anforderungen der Verordnung (Art. 11 Abs. 1 UAbs. 2, EG 46 KOM-E). Dies erstreckt sich jedoch nur auf Behörden und sonstige staatliche Stellen (v. a. Art. 64 Abs. 3 KOM-E, s. a. EG 79 KOM-E). Die Marktüberwachungsbehörden haben nach Art. 64 Abs. 1 KOM-E sogar uneingeschränkten Zugang zu den von Anbietern genutzten Trainings-, Validierungs- und Testdatensätzen, auch über Anwendungsprogrammierschnittstellen (API) oder sonstige für den Fernzugriff geeignete technische Mittel und Instrumente. Demgegenüber findet sich an keiner Stelle eine Regelung dazu, ob die Adressaten von Hochrisiko-KI-Systemen Zugang zur Dokumentation erhalten, um ihre Verbraucherschutzrechte zu wahren. Die Regelungen zur Dokumentation teilen insoweit die allgemeine Schieflage des KOM-E, der sehr stark auf eine behördlich überwachte Regulierung der KI-Systeme setzt, dabei jedoch Verbraucherinnen und Verbraucher als Akteure völlig außen vor lässt.

IV. Regulierungsbedarf jenseits des Datenschutzrechts?

Als Zwischenergebnis lässt sich festhalten, dass das geltende Datenschutzrecht durchaus Instrumente enthält, um die Rechte von betroffenen Personen hinsichtlich der Verwendung ihrer personenbezogenen Daten zum Training von KI-Systemen sowie hinsichtlich des späteren Einsatzes von mit personenbezogenen Daten trainierten KI-Systemen zu wahren. Allerdings bleiben diese Instrumente aufgrund des Erfordernisses des Personenbezugs fragmentarisch und enthalten teilweise deutliche tatbestandliche Einschränkungen bzw. Unklarheiten. Dies führt zur Frage, ob der Gesetzgeber tätig werden sollte.

1. Unterlassungsansprüche gegen unzulänglich trainierte KI?

Fraglich ist insbesondere, ob die Vorgaben aus Art. 10 Abs. 3 S. 1 KOM-E aus der Perspektive von Verbraucherinnen und Verbrauchern fruchtbar gemacht werden können. Mit anderen Worten: Wie können diese sich dagegen wehren, zu Adressaten von KI-Systemen zu werden, wenn die Trainings-, Validierungs- und Testdatensätze nicht relevant, repräsentativ, fehlerfrei und vollständig waren?

Das geltende Datenschutzrecht erweist sich dabei erneut als unzulänglich. Sofern das System mit eigenen, unrichtigen personenbezogenen Daten der Adressaten trainiert, validiert oder getestet wurde, besteht ein Berichtigungsanspruch

nach Art. 16 DSGVO. Dies dürften jedoch Sonderfälle sein. Unrichtige Daten über andere Personen, unrichtige anonyme Daten und unrichtige Sachdaten werden nicht erfasst. Dasselbe gilt für irrelevante und nicht repräsentative Daten: Der Berichtigungsanspruch ist ein Individualrecht, das vor unzutreffenden Fremdbildern schützen soll; selbst im Falle der Verwendung eigener personenbezogener Daten geht dies aber nicht so weit zu verhindern, dass richtige Daten für Zwecke verwendet werden, für die sie nicht relevant sind. Zwar könnte man insoweit den übergeordneten Grundsatz der Datenrichtigkeit (Art. 5 Abs. 1 lit. d DSGVO) oder den Grundsatz von Treu und Glauben (Art. 5 Abs. 1 lit. a DSGVO) in Erwägung ziehen. Allerdings dürfte es ausgeschlossen sein, aus diesen einen allgemeinen subjektiven Anspruch auf Unterlassung der Anwendung eines unzulänglich trainierten KI-Systems abzuleiten.[54]

Je nach der späteren Verwendungsumgebung kann sich ein solcher Anspruch allerdings auch aus dem Datenschutzrecht oder aus anderen Rechtsgrundlagen ergeben. Wenn ein unzulänglich trainiertes KI-System nunmehr unrichtige personenbezogene Daten produziert, so greift Art. 16 DSGVO. Außerhalb des Datenschutzrechts werden in vielen Fällen Mängelgewährleistungsansprüche, vertraglichen Nebenpflichten, das Produkt- oder Produzentenhaftungsrecht oder Ansprüche aus dem Staatshaftungsrecht[55] einschlägig sein. Ob sich insoweit aus der Perspektive von Verbraucherinnen und Verbrauchern relevante Lücken bei der Durchsetzung ihrer Rechte ergeben, bedürfte einer weitergehenden Analyse.

Prima facie dürften die eigentlichen Probleme weniger in der Verfügbarkeit der jeweiligen Abwehransprüche als vielmehr auf der Beweisebene liegen. Die vielfach diskutierten „Blackbox"-Probleme beim Einsatz von KI greifen auch hier: Je elaborierter die Systeme werden und je mehr sie in den verschiedenen Phasen der Entwicklung, der Konfiguration und des späteren Einsatzes trainiert werden, desto schwerer werden Verbraucherinnen und Verbraucher den Nachweis erbringen können, dass im Einzelfall unrichtige oder unangemessener Ergebnisse produziert werden oder gar dass diese Ergebnisse auf die Verwendung von Trainings-, Validierungs- und Testdatensätze zurückgehen, die nicht relevant, repräsentativ, fehlerfrei und vollständig waren. Dies ist ein systemisches Problem, da die (Un-)Fähigkeit von KI-Systemen, Entscheidungsgrundlagen und Entscheidungskriterien dem Benutzer zu erklären, also für ihn nachvollziehbar zu machen, sich oftmals auch auf den Nutzer erstreckt.[56] Der EuGH hat unlängst hervorgehoben,

---

[54] S. zur unklaren Reichweite des Grundsatzes der Richtigkeit bei Anwendung auf KI *Roßnagel*, in Simitis/Hornung/Spiecker gen. Döhmann (Fn. 34), Art. 5 Rn. 148 f., *Hacker*, ZGE 2020, 239, 245; s. ferner die Überlegungen bei *Hoeren*, MMR 2016, 8.

[55] Speziell zu dieser Frage und den Herausforderungen im Bereich von Kausalität und Verschulden s. *Roth-Isigkeit*, AöR 145 (2020), 321; *Martini/Ruschemeier/Hain*, VerwArch 112 (2021), 1.

[56] S. zu diesem zentralen Problem z. B. *Sudmann*, Digital Culture & Society 2018, 181; *Wischmeyer*, AöR 143 (2018), 1, 42 ff.; *Martini*, Blackbox Algorithmus, 2019, 28 ff.; *Guckelberger*, Öffentliche Verwaltung im Zeitalter der Digitalisierung, 2019, 520 ff.; *Malgieri*, Computer Law & Security Review 35 (2019) 105327; *Käde/von Maltzan*, CR 2020, 66.

dass KI-Systeme, die an einer mangelnden Nachvollziehbarkeit leiden und es dadurch unmöglich machen, den Grund für einen Grundrechtseingriff zu erkennen (konkret: einen Treffer in einer Passenger Name Records (PNR)-Datenbank), mit dem Recht auf einen wirksamen gerichtlichen Rechtsbehelf nach Art. 47 GRCh in Konflikt kommen können.[57] Ansätze zu einer technischen Lösung im Sinne einer „Explainable AI"[58] stellen derzeit eine der zentralen rechtlichen Forderungen und zugleich eine der größten technischen Herausforderung der KI-Forschung dar. Perspektivisch wird hier die Frage zu beantworten sein, ob diesen Problemen durch erweiterte Dokumentationspflichten (wie im KOM-E, aber erweitert um eine Zugänglichkeit für die Adressaten der KI-Systeme), Beweislasterleichterungen, Ansprüche auf Begründung und Erklärung[59] oder vorverlagerte Abwehransprüche begegnet werden muss.

2. Regulierungsbedarf für anonyme und synthetische Trainingsdaten?

Im Anschluss an diese Überlegungen stellt sich die Frage, ob der europäische oder der nationale Gesetzgeber den Umgang mit Trainingsdaten umfassend regulieren sollte. Dies würde nicht nur personenbezogene Daten, sondern auch anonyme Daten sowie synthetische Daten erfassen, also solche Trainingsdaten, die speziell für das Training generiert werden und beispielsweise personenbezogene Daten simulieren.[60]

a) Fehlende Eignung des Anwendungsbereichs und der Schutzinstrumente des Datenschutzrechts

Die deutlichste Einschränkung des Datenschutzrechts hinsichtlich seiner Eignung zum Schutz von Verbraucherinnen und Verbrauchern beim Einsatz von KI-Sys-

---

[57] EuGH, Urt. v. 21.6.2022 – C-817/19 – PNR-Daten, Rn. 194 f. = EuZW 2022, 706.

[58] S. aus verschiedenen Perspektiven *O'Hara*, Computer Law & Security Review 39 (2020), 105474; *Rohlfing u. a.*, Explanation as a social practice, IEEE Transactions on Cognitive and Developmental Systems, DOI: 10.1109/TCDS.2020.3044366; zu Umsetzungsmöglichkeiten z. B. *Waltl/Vogl*, DuD 2018, 613; *Käde/von Maltzan*, CR 2020, 66, 69 ff.; *Körner*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 2.4; *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 28 (2020), 415; *Bibal/Lognoul/de Streel/Frénay*, Artificial Intelligence and Law 29 (2021), 149 ff.

[59] Für ein subjektives „Recht auf Erklärung" nach der DSGVO *Vogel*, Künstliche Intelligenz und Datenschutz, 2021, 172 ff.; Vorschlag für eine umfassende „reviewability" bei *Cobbe/Singh*, Computer Law & Security Review 39 (2020), 105475; s. auch die Überlegungen bei *Busch*, Algorithmic Accountability, ABIDA Gutachten, 2018, 56 ff.; *Martini*, Blackbox Algorithmus, 2019, 176 ff.; *Sesing*, MMR 2021, 288; zu Transparenzfragen auch *Wischmeyer*, in: Wischmeyer/Rademacher (Hrsg.), Regulating Artificial Intelligence, 2020, 75 ff.; Vorschläge de lege ferenda bei *Martini*, Blackbox Algorithmus, 2019, 340 ff.

[60] *Meents*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.8 Rn. 45 ff.; *Kaulartz*, ebd., Kap. 8.9 Rn. 22 ff.; ausführliche rechtliche Bewertung bei *Raji*, DuD 2021, 303; Lösungswege für datenschutzkonformes Training auch bei *Boenisch*, DuD 2021, 448; *Stock/Petersen/Behrendt/Federrath/Kreutzburg*, Informatik Spektrum 45 (2022), 137.

temen ist die Beschränkung auf personenbezogene Daten. Das Datenschutzrecht schützt zwar nicht nur das Grundrecht auf Schutz personenbezogener Daten, sondern – schon ausweislich Art. 1 Abs. 2 DSGVO – auch weitere Grundrechte und Grundfreiheiten natürlicher Personen. Dies bezieht sich insbesondere auf den Schutz vor Diskriminierungen (Art. 20 ff. GRCh).[61] Dies erweitert den sachlichen Anwendungsbereich nach Art. 2 Abs. 1 DSGVO jedoch nicht, sodass die Auswirkungen der Verwendung ungeeigneter oder verzerrender anonymer oder sachbezogener Trainingsdaten nicht erfasst werden.

Hinzu kommen die soeben beschriebenen Probleme einer Ergebniskontrolle: Diese birgt das Risiko, zu spät zu kommen und an Beweisproblemen zu scheitern. Etliche bekannt gewordene Fälle von verzerrenden Ergebnissen, Diskriminierungen und vergleichbaren Effekten deuten außerdem darauf hin, dass die intrinsische Motivation der Anbieter und Nutzer von KI-Systemen zur Verwendung hochwertiger Trainingsdaten[62] nicht hinreichend ist, um derartige negative Wirkungen zu vermeiden. Führt man sich schließlich die besondere Bedeutung der Trainingsdaten für die Auswirkungen von KI-Systemen vor Augen, so spricht viel dafür, der „Algorithmenkontrolle"[63] eine Datenkontrolle zumindest an die Seite zu stellen.[64]

Das Datenschutzrecht ist für eine solche Kontrolle noch aus weiteren Gründen nicht hinreichend geeignet. Selbst wenn sein sachlicher Anwendungsbereich eröffnet ist, ist es typischerweise blind für etwaige Folgewirkungen der Zulässigkeit oder Unzulässigkeit einer Datenverarbeitung für Dritte. Mit Blick auf das Ziel der Entwicklung hochwertiger KI-Systeme kann das Datenschutzrecht damit einerseits zu eng, andererseits aber auch zu weit sein:

Einerseits ist es für die Frage der Zulässigkeit einer Datenverarbeitung nach Art. 6 Abs. 1 DSGVO typischerweise irrelevant, ob die Verarbeitung negative Folgen für Dritte mit sich bringt. Dies ist besonders deutlich bei der Einwilligung, gilt aber auch für die Interessenabwägung nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO, bei der zwar auf Seiten des Verantwortlichen auch die berechtigten Interessen eines Dritten in Anschlag gebracht werden können, nicht aber auf Seiten der betroffenen Person. Mit anderen Worten spielt es für die Zulässigkeit der Trainingsdatenverarbeitung auf Basis einer Einwilligung oder der Interessenabwägung keine Rolle, ob hierdurch negative Auswirkungen auf andere zu erwarten sind.

Andererseits kann es dazu kommen, dass Trainingsdaten aus individuellen, in der betroffenen Person liegenden Gründen nicht verarbeitet werden dürfen (bei-

---

[61] S. *Hornung/Spiecker gen. Döhmann*, in: Simitis/Hornung/Spiecker gen. Döhmann (Fn. 34), Art. 1 Rn. 31 f., 36 ff.

[62] S. o. unter II. 2.

[63] S. *Martini*, DVBl 2014, 1481 (1488).

[64] Zur Notwendigkeit s. insoweit schon Europäische Kommission, Weißbuch zur Künstlichen Intelligenz, COM(2020) 65 final, 22 f.; *Hacker*, NJW 2020, 2142, 2144 f.; *ders.*, ZGE 2020, 239, 242 ff., 259 ff.

spielsweise wegen möglicher negativer Folgen für sie, Art. 6 Abs. 4 lit. d DSGVO). Auch dabei spielt es keine Rolle, ob durch diese Unzulässigkeit und das dadurch verursachte Fehlen eines einzelnen Trainingsdatensatzes in einer entsprechenden Datensammlung beispielsweise Verzerrungseffekte in den Trainingsdaten eintreten und damit wiederum Dritte Nachteile erleiden.

Schlussendlich können auch die typischen Schutzinstrumente des Datenschutzrechts Probleme für den Einsatz von Trainingsdaten mit sich bringen. Schon die Pseudonymisierung, insbesondere aber die Anonymisierung kann die Qualität der Trainingsdaten verschlechtern, wenn sie ernsthaft betrieben werden. Denn angesichts des sehr breiten Begriffs des personenbezogenen Datums[65] muss für eine echte Anonymisierung erheblich mehr unternommen werden, als lediglich den Namen zu entfernen. Zusätzlich ist die Entfernung solcher Teile der verbleibenden Daten erforderlich, die es mit entsprechendem Zusatzaufwand ermöglichen, die betroffene Person zu ermitteln.[66] Für diese Frage sind nach EG 26 S. 3 DSGVO alle Mittel zu berücksichtigen, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren. Hierfür spielen nach EG 26 S. 4 DSGVO alle objektiven Faktoren eine Rolle, insbesondere die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, jeweils bezogen auf die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen. Dies kann eine Einzelfallbetrachtung erfordern und je nach technologischem Fortschritt auch ein „schleichendes" Eintreten des Personenbezugs implizieren.[67] Eine echte Anonymisierung kann damit zwar zum einen dem Verantwortlichen die datenschutzrechtskonforme Verwendung der Daten zum Training von KI-Systemen ermöglichen.[68] Will der Verantwortliche allerdings vollständig anonymisieren und damit auch das Risiko ausschließen, dass die trainierten Modelle nachträglich die personenbezogenen Daten preisgeben,[69] wird

---

[65] Im Streit zwischen den sogenannten relativen und absoluten Begriffen des Personenbezugs (dazu im Überblick *Hofmann/Johannes*, ZD 2017, 221; ausführlich *Schmidt-Holtmann*, Der Schutz der IP-Adresse im deutschen und europäischen Datenschutzrecht, 2014; *Haase*, Datenschutzrechtliche Fragen des Personenbezugs, 2015) hat sich der EuGH zwar formal für den relativen Begriff entschieden, der auf die Identifizierbarkeit für den konkreten Verantwortlichen abstellt (s. EuGH, Urt. v. 19.10.2016, Rs. C-582/14, NJW 2016, 3579 – Breyer). Zugleich sind die Kriterien, die das Gericht jedenfalls für den Personenbezug der IP-Adresse angibt, so weit gefasst, dass das Ergebnis zumindest in diesem Fall dem eines absoluten Begriffs des Personenbezugs ähnelt.

[66] S. zu den Anforderungen an die Anonymisierung insoweit *Hansen*, in: Simitis/Hornung/Spiecker gen. Döhmann (Fn. 34), Art. 4 Nr. 5 Rn. 50 ff.; speziell für Trainingsdaten *Niemann/Kevekordes*, CR 2020, 17, 19; allgemeiner für KI-Systeme *Vogel*, Künstliche Intelligenz und Datenschutz, 2022, 216 ff.

[67] Dazu *Hornung/Wagner*, CR 2019, 565.

[68] S. *Roßnagel/Geminn*, ZD 2021, 487; dort auch zur Notwendigkeit einer Vorsorgeregelung zur Verhinderung der Herstellung des Personenbezugs. Zur Zulässigkeit der Anonymisierung als datenschutzrechtliche Verarbeitung s. *Hornung/Wagner*, ZD 2020, 223 m. w. N.

[69] S. dazu *Kaulartz*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.9 Rn. 9 ff.

er im Zweifel substantielle Teile der Daten entfernen und die damit verbundenen Einschränkungen der Qualität der Trainingsdaten hinnehmen müssen.

Ähnliche Probleme können sich durch die Beachtung der Grundsätze der Verarbeitung nach Art. 5 DSGVO einstellen.[70] So verlangt beispielsweise der Grundsatz der Datenminimierung, dass personenbezogene Daten dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein müssen (Art. 5 Abs. 1 lit. c DSGVO).[71] Der Maßstab für die Notwendigkeit kann jedoch nicht der des späteren Einsatzzweckes eines KI-Systems sein. Denn um solche Systeme zu trainieren, bedarf es einer Negativkontrolle, die verhindert, dass ein System falschpositive Ergebnisse liefert. Mit anderen Worten: Man kann ein KI-System nicht darauf trainieren, PKWs zu erkennen, in dem man es ausschließlich mit Bildern von PKWs trainiert. Derartige Notwendigkeiten führen dazu, dass sowohl in der Trainingsphase als auch für eine spätere Qualitäts- und Ergebniskontrolle gerade solche Daten gebraucht werden, die auf den ersten Blick wegen Art. 5 Abs. 1 lit. c DSGVO nicht verarbeitet werden dürften.

Ein weiteres Beispiel für einen vergleichbaren Effekt ist der Grundsatz der Speicherbegrenzung nach Art. 5 Abs. 1 lit. e DSGVO. Danach müssen personenbezogene Daten anonymisiert oder gelöscht werden, wenn sie für die Zwecke, für die sie verarbeitet werden, nicht mehr erforderlich sind. Dies könnte auf den ersten Blick dafür sprechen, nach Abschluss des Trainings eines KI-Systems die verwendeten Trainingsdaten zu löschen. Allerdings würde dies einen späteren Vergleich mit der Leistungsfähigkeit eines anderen KI-Systems erschweren, weil dies die Verwendung identischer Trainingsdaten erfordern kann. Auch für Dokumentationen oder spätere Haftungsfragen kann sich eine fortdauernde Relevanz der Daten ergeben, die einer Löschung entgegensteht.

b) Vorschläge im KOM-E

Mit Art. 10 KOM-E unternimmt die Kommission den Versuch eines umfassenden regulatorischen Zugriffs auf die Trainings-, Validierungs- und Testdatensätze von Hochrisiko-KI-Systemen. Dabei wird die größte Einschränkung des Datenschutzrechts beseitigt, da nicht nur personenbezogene Daten von den Qualitätskriterien erfasst sind. Dieses Modell – das auch der schleswig-holsteinische Gesetzgeber in § 8 Abs. 3 ITEG SH gewählt hat – ist auch aus Verbraucherschutzsicht zu begrüßen.

Ähnliches gilt für den Ansatz, losgelöst von den Grundsätzen der Verarbeitung in Art. 5 DSGVO selbstständige Qualitätskriterien zu regulieren, die den spezifischen Erfordernissen des Trainings von KI-Systemen entsprechen. Dabei gibt es

---

[70] S. allgemein zu KI und den Grundsätzen der Datenverarbeitung *Paal*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.7 Rn. 4 ff.; s. a. *Vogel*, Künstliche Intelligenz und Datenschutz, 2022, 86 ff.

[71] Forderung der Anwendung auf KI-Trainingsdaten z. B. im Gutachten der *Datenethikkommission*, 2019, 120; zur Umsetzung *Meents*, in: Kaulartz/Braegelmann (Fn. 4), Kap. 8.8 Rn. 9 ff.

im Detail zwar noch Verbesserungsbedarf.[72] Es ist aber erkennbar, dass beispielsweise das Kriterium der Relevanz eines Datensatzes für das Training, die Validierung oder das Testen eines KI-Systems (Art. 10 Abs. 3 S. 1 KOM-E) deutlich besser geeignet ist als das Kriterium, ob es im datenschutzrechtlichen Sinne auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt ist (Art. 5 Abs. 1 lit. c DSGVO).

Noch offen ist insoweit allerdings die Frage des erforderlichen Aufwands aufseiten der Anbieter. Die Anforderungen in Art. 10 Abs. 2 bis Abs. 4 KOM-E sind (ebenso wie die in § 8 Abs. 3 ITEG SH) bemerkenswert kategorisch formuliert. Nimmt man den Wortlaut ernst, so dürften etliche Anbieter in erhebliche Probleme kommen, denn gerade bei innovativen KI-Systemen kann es vorkommen, dass sich erst im Nachgang Unzulänglichkeiten der Trainings-, Validierungs- und Testdatensätze herausstellen. Auch bei gut verstandenen KI-Systemen dürfte es selbst mit erheblichem Aufwand unmöglich sein, in jedem Einzelfall zu 100 % relevante, repräsentative, fehlerfreie und vollständige (Art. 10 Abs. 3 S. 1 KOM-E) oder zu 100 % nicht-diskriminierende, integre, objektive und valide (§ 8 Abs. 3 ITEG SH) Daten zu verwenden.[73] Eine wichtige Frage wird deshalb sein, nach welchem zeitlichen Horizont sich die Beurteilung richtet. Außerdem wird man nicht umhinkommen, den zumutbaren Aufwand für Anbieter abzugrenzen, den sie im Rahmen von Art. 10 Abs. 2 bis Abs. 4 KOM-E betreiben müssen. Dazu bedarf es einer Diskussion darum, wie viele Fehler in den Datensätzen und welche Fehlerverteilung (z. B. auf unterschiedliche Adressaten oder Adressatengruppen) noch tolerabel sind. Die Formulierung von Art. 10 Abs. 3 und Abs. 4 KOM-E spricht dafür, dass insoweit im Bereich von Hochrisiko-KI-Systemen sehr hohe Maßstäbe anzulegen sind.

In bestimmten Fällen wird man die Anforderungen der Relevanz, Repräsentativität, Fehlerfreiheit und Vollständigkeit außerdem auf den konkreten Einsatzzweck und das konkrete Trainingsmodell beziehen und insoweit einschränkend interpretieren müssen. Wenn ein KI-System beispielsweise auf die Erkennung sehr seltener Ereignisse (schwere Verkehrsunfälle o. ä.) trainiert werden soll, so kann es dazu kommen, dass die verwendeten Trainingsdaten nur sehr eingeschränkt repräsentativ und in keiner Weise vollständig sind. Sofern man sich hierüber allerdings beim Training bewusst ist, können mit geeigneten KI-Algorithmen, die dies berücksichtigt, doch gute KI-Modelle trainiert werden. Sofern dies erfolgt und die verbleibenden Einschränkungen der Aussagekraft der Trainingsergebnisse transparent gemacht werden, sollte dies nicht als Verstoß gegen Art. 10 Abs. 3 KOM-E verstanden werden.

---

[72] S. dazu den Beitrag von *Hacker* in diesem Band.

[73] S. *Hacker*, ZGE 2020, 239, 265 f.; kritisch zum Entwurf aus diesem Grund *Ebers/Hoch/Rosenkranz/Ruschemeier/Steinrötter*, RDi 2021, 528, 533.

Schließlich ist in diesem Zusammenhang der Vorschlag für die Zulässigkeit der datenschutzrechtlichen Zweckänderung im Rahmen der Verarbeitungsbefugnis nach Art. 10 Abs. 5 KOM-E erneut zu würdigen. Diese knüpft die Kommission einerseits gerade nicht an eine Einwilligung, sondern an die Erforderlichkeit für die Beobachtung, Erkennung und Korrektur von Verzerrungen im Zusammenhang mit Hochrisiko-KI-Systemen. Dies ist funktional angemessen, weil sonst das Risiko bestünde, dass Verzerrungseffekte nicht erkannt oder sogar verstärkt würden, wenn einzelne betroffene Personen die Einwilligung erteilen, andere sie verweigern. Andererseits statuiert der KOM-E einen verschärften Erforderlichkeitsmaßstab und verbindet diesen mit der Vorgabe „modernster" technischer und organisatorischer Sicherheit- und Datenschutzmaßnahmen.[74] Dies erscheint als ein angemessener Kompromiss zwischen den Notwendigkeiten des Trainings von KI-Systemen zur Vermeidung von Diskriminierungen einerseits, den berechtigten Datenschutzinteressen von Verbraucherinnen und Verbrauchern andererseits.

Begleitet werden die Anforderungen in Art. 10 KOM-E durch Transparenzvorgaben, die über die des Datenschutzrechts hinausgehen.[75] Relevante Informationen über die verwendeten Trainings-, Validierungs- und Testdatensätze unter Berücksichtigung der Zweckbestimmung des KI-Systems müssen dem Nutzer nach Art. 13 Abs. 2 i V m Abs. 3 lit. b KOM-E bereitgestellt werden. Erforderlich sind präzise, vollständige, korrekte und eindeutige Informationen in einer für die Nutzer relevanten, barrierefrei zugänglichen und verständlichen Form. Freilich ist der Nutzer eben nicht die Verbraucherin oder der Verbraucher, sondern nach Art. 3 Nr. 4 KOM-E der Verwender des KI-Systems,[76] und über eine Weitergabe der Informationen an die betroffenen Verbraucherinnen und Verbraucher schweigt der Entwurf. Für den Umgang mit der nach § 8 Abs. 4 ITEG SH für öffentliche Stellen verpflichtenden Dokumentation gibt es sogar gar keine Regelungen hinsichtlich der weiteren Verwendung und etwaiger Auskunftsrechte Dritter; auch die Verordnungsermächtigung in § 11 Abs. 1 S. 2 Nr. 3 ITEG SH bezieht sich nur auf die Dokumentations- und Protokollierungspflichten selbst.

Dagegen enthält Art. 52 KOM-E Transparenzpflichten, die gegenüber den von KI-Systemen betroffenen natürlichen Personen zu erfüllen sind. Diese erfassen insbesondere den Umstand, dass mit einem KI-System interagiert wird (Art. 52 Abs. 1 KOM-E) sowie Fälle der Emotionserkennung und Biometrie (Abs. 2) wie auch von Deepfakes (Abs. 3). Die Pflichten beziehen sich aber nicht auf die Transparenz hinsichtlich der verwendeten Trainings-, Validierungs- und Testdatensätze.

---

[74] S. o. unter III. 2. c) aa).
[75] S. *Spindler*, CR 2021, 361, 368.
[76] S. o. unter II. 2.

V. Ausblick

Im Ergebnis ist der Vorschlag der Kommission zum Umgang mit Trainings-, Validierungs- und Testdatensätzen bei allen Diskussionspunkten im Detail ein Schritt in die richtige Richtung. Er löst die Regulierung von Trainingsdaten aus dem datenschutzrechtlichen Korsett, ohne Brüche mit diesem zu verursachen und ohne das Kind mit dem Bade auszuschütten, indem für das Ziel einer Führungsrolle der Europäischen Union bei der Entwicklung von KI-Systemen[77] personenbezogene Daten von Verbraucherinnen und Verbrauchern völlig freigegeben würden.

Angesichts der für Verstöße gegen Art. 10 KOM-E geltenden Sanktionen steht auch zu erwarten, dass die Norm in der Praxis beachtet werden wird. Die Nichtkonformität eines KI-Systems mit den Anforderungen aus Art. 10 KOM-E soll nach Art. 71 Abs. 3 lit. b KOM-E mit Geldbußen von bis zu 30 Mio. € oder – im Falle von Unternehmen – von bis zu 6 % des gesamten weltweiten Jahresumsatzes des vorangegangenen Geschäftsjahres belegt werden können, je nachdem, welcher Betrag höher ist. Für Organe, Einrichtungen und sonstige Stellen der Union beträgt die mögliche Geldbuße nach Art. 72 Abs. 2 lit. b KOM-E immerhin noch 500.000 €.

Allerdings verbleiben zwei grundsätzliche Probleme des Kommissionsentwurfs, nämlich zum einen der fehlende Fokus auf Verbraucherinnen und Verbraucher und zum anderen die Beschränkung von Art. 10 KOM-E auf Hochrisiko-KI-Systeme.

Der erste Punkt ist allgemeiner Natur. Der Entwurf nimmt deutlich zu wenig die Perspektive der Adressaten von KI-Systemen ein. Der starke Fokus auf eine behördliche Durchsetzung der materiellrechtlichen Anforderungen im gesamten KOM-E muss zumindest ergänzt werden um eine auch prozedural gestärkte Rechtsposition von Verbraucherinnen und Verbrauchern. Es ist bezeichnend, dass Verbraucherschutz im Entwurf nur sehr allgemein genannt wird (S. 4, 13, 15, EG 28). Als Personen werden Verbraucherinnen und Verbraucher ausschließlich im Finanzbogen zum KOM-E erwähnt, und auch hier nur mit dem lapidaren Satz, sie „sollten davon profitieren, dass das Risiko von Verletzungen ihrer Sicherheit oder ihrer Grundrechte eingedämmt wird".[78] Grundrechtsverletzungen sind hingegen nicht nur etwas, das der europäische Gesetzgeber einzudämmen hat, sondern sollten auch sekundärrechtlich zu individuellen Rechtsschutzmöglichkeiten führen.[79]

---

[77] COM(2021) 206 final, S. 1.

[78] COM(2021) 206 final, S. 104.

[79] S. zu Ansprüchen der Betroffenen als Teil eines Regulierungsrahmens für Trainingsdaten *Hacker*, ZGE 2020, 239, 268 ff.; fehlende Individualrechte werden auch kritisiert von *Ebers/Hoch/Rosenkranz/Ruschemeier/Steinrötter*, RDi 2021, 528, 537.

Es ist zwar durchaus plausibel, dass die materiellen Anforderungen an KI-Systeme auch mit den bestehenden Rechtsschutzinstrumenten des Verbraucherschutzrechts aktiviert werden können. Denn die Festlegung von Standards beispielsweise zum Einsatz qualitativ hochwertiger Trainings-, Validierungs- und Testdatensätze in Art. 10 KOM-E werden auch Auswirkungen auf Produkteigenschaften oder Verkehrserwartungen haben.[80] Das aktuelle Rechtsetzungsverfahren bietet aber die Chance, Verbraucherinnen und Verbrauchern den Zugriff auf zumindest einige der neuen Governance-Instrumente zu geben. Hiervon sollte der europäische Gesetzgeber Gebrauch machen.

Der zweite Punkt, die Beschränkung von Art. 10 KOM-E auf Hochrisiko-KI-Systeme, wurde bereits hinsichtlich der unklaren Auswirkungen auf die Verarbeitungsbefugnis in Art. 10 Abs. 5 KOM-E thematisiert.[81] Hinsichtlich der Anforderungen an Trainings-, Validierungs- und Testdatensätze in Art. 10 Abs. 2 bis Abs. 4 KOM-E ist der limitierte Regelungsansatz auf den ersten Blick noch schwerer verständlich – es wäre nicht zu rechtfertigen (und ist von der Kommission sicher nicht beabsichtigt), den Anbietern und Nutzern von „normalen" KI-Systemen im Umkehrschluss zu Art. 10 Abs. 3 KOM-E die Verwendung irrelevanter, nicht repräsentativer, falscher und unvollständiger Trainingsdaten zu gestatten.

Erklärbar ist die Beschränkung des Anwendungsbereichs durch den abgestuften Regulierungsansatz des KOM-E, der lediglich als besonders riskant bewertete KI-Systeme strengeren materiellrechtlichen und erheblichen verfahrensrechtlichen Anforderungen unterwerfen will. KI-Systeme, die unterhalb dieser Schwelle bleiben, sollen aus Verhältnismäßigkeitsgründen nicht mit bürokratischen Vorgaben überfrachtet werden. Zumindest bei Art. 10 KOM-E wäre insoweit allerdings eine Trennung zwischen materiellrechtlichen Vorgaben einerseits, verfahrensrechtlichen Anforderungen und Sanktionen andererseits angezeigt. Insbesondere die sehr hohen Bußgelder in Art. 71 Abs. 3 lit. b KOM-E müssten für normale KI-Systeme entfallen oder erheblich abgesenkt werden. Auch bei der Frage, wie streng die Anforderungen an die Qualität der Trainings-, Validierungs- und Testdatensätze zu fassen sind, wird man Abstriche machen müssen. Denn wenn sogar bei Hochrisiko-KI-Systemen die Vorgaben der Art. 10 Abs. 2 bis Abs. 4 KOM-E nicht zu 100 % erfüllbar sind, kann dies von den Anbietern anderer KI-Systeme nur in abgestufter Form erwartet werden.[82] Das ändert aber nichts daran, dass aus Verbraucherschutzsicht eine Regelung zu den materiellen Anforderungen an

---

[80] S. zum Charakter der Bestimmungen des KOM-E als Schutzgesetze i. S. v. § 823 Abs. 2 BGB *Grützmacher*, CR 2021, 433; dies wird dort bejaht (437 ff., speziell zu Art. 10 KOM-E 439 f.); in diese Richtung auch *Spindler*, CR 2021, 361, 362; zu haftungsrechtlichen Fragen der Qualität von Trainingsdaten auch *Hacker*, ZGE 2020, 239, 249 ff.; *Zech*, NJW 2022, 502; zum Staatshaftungsrecht s. Fn. 55.

[81] S. o. unter III. 2. c) aa).

[82] S. a. *Hacker*, ZGE 2020, 239, 267.

Trainings-, Validierungs- und Testdatensätze für alle KI-Systeme begrüßenswert wäre, weil so auch klargestellt würde, dass Verbraucherinnen und Verbraucher eine berechtigte Verkehrserwartung dahingehend haben, dass sämtliche KI-Systeme in der Praxis entsprechende Qualitätsanforderungen einhalten.

Konfliktlinien: Geheimhaltungsinteressen vs. Transparenz von ADM-Systemen

Ruth Janal

I. Einführung

Entscheidungen oder Empfehlungen, die von automatisierten Entscheidungssystemen (ADM-Systeme) ausgesprochen werden, sind für die betroffenen Personen und die Öffentlichkeit oftmals schwer nachvollziehbar. Die Gründe sind vielfältig: Die den Entscheidungen zugrunde liegenden Datensätze sind gegebenenfalls nicht frei verfügbar, die eingesetzte Software ist proprietär oder die auf *deep learning* beruhenden Entscheidungsprozesse sind für Menschen per se nicht erklärlich. Doch die Forderungen nach einer besseren Nachvollziehbarkeit und Transparenz algorithmischer Prozesse werden lauter. Mit der Veröffentlichung der sog. „Facebook Files" durch das Wall Street Journal[1] sowie der Aussage der Facebook-Whistleblowerin *Frances Haugen* vor dem US-Senat[2] haben diese Forderungen einen vorläufigen Höhepunkt gefunden.

Gegenwärtig erfolgt die Darstellung der Funktionsweise und Funktionsfähigkeit von automatisierten Entscheidungssystemen oftmals durch die Marketing-Abteilungen derjenigen Unternehmen, welche die Systeme entwickelt haben. Kritische interne Forschung ist entweder von vorneherein nicht zur Veröffentlichung vorgesehen, oder die Publikation unterliegt intensiven Beschränkungen.[3] Externe Forscher:innen sind auf die freiwillige Eröffnung des System- bzw. Datenzugangs angewiesen und deshalb nicht vollkommen unabhängig.[4] Auch bestehen Zweifel, ob die Daten, welche externen Forschern zur Verfügung gestellt werden, ein voll-

---

[1] *Horwitz* et al., the facebook files, https://www.wsj.com/articles/the-facebook-files-11631713039.

[2] *Paul*, Facebook whistleblower hearing: Frances Haugen calls for more regulation of tech giant – as it happened, 5.10.2021, https://www.theguardian.com/technology/live/2021/oct/05/facebook-hearing-whistleblower-frances-haugen-testifies-us-senate-latest-news.

[3] *Coulter*, Google's AI researchers say their output is being slowed by lawyers after a string of high-level exits: ‚Getting published really is a nightmare right now', 22.10.2021, https://www.businessinsider.com/google-ethical-ai-timnit-gebru-2021-10?op=1; *Vincent*, Google is poisoning its reputation with AI researchers, 13.4.2021, https://www.theverge.com/2021/4/13/22370158/google-ai-ethics-timnit-gebru-margaret-mitchell-firing-reputation.

[4] Siehe näher *Lapowsky*, Platforms vs. PhDs: How tech giants court and crush the people who study them, 19.3.2021, https://www.protocol.com/nyu-facebook-researchers-scraping.

ständiges Bild zeichnen.[5] Ableitungen aus (begrenzten) Datenanalyseinstrumenten, die der Allgemeinheit von Unternehmensseite zur Verfügung gestellt werden, lassen sich von Unternehmensseite bestreiten, ohne dass eine Überprüfungsmöglichkeit bestünde.[6] Zudem werden unautorisierte externe Analysen gegebenenfalls durch das Androhen rechtlicher Schritte[7] oder das Ergreifen technischer Maßnahmen, die einen Datenzugriff erschweren,[8] unterbunden. Mangels externer Validierung besteht das Risiko des Einsatzes nicht ausreichend getesteter Entscheidungssysteme in allen gesellschaftlichen Bereichen, beispielsweise im Bildungssektor, in der Medizin oder bei der Kreditvergabe.

Der gesellschaftlichen Forderung nach einer Transparenz automatisierter Entscheidungssysteme werden von Seiten der Entwickler und Betreiber oftmals zwei Argumente entgegengehalten: Der Schutz personenbezogener Daten und der Schutz von Geschäftsgeheimnissen. Daneben steht auch das Ziel, eine Manipulation der Systeme zu verhindern. Mit Blick auf dieses Spannungsfeld erörtere ich im vorliegenden Beitrag zunächst, welche Geheimhaltungsinteressen bestehen und inwiefern diesen rechtlicher Schutz zukommt (II.). In einem zweiten Schritt stelle ich die gegenwärtig verankerten gesetzlichen Offenbarungspflichten vor, die eine Transparenz algorithmischer Entscheidungsprozesse ermöglichen sollen (III.). Im Anschluss erörtere ich, ob Versuche von Wissenschaftler:innen und Zivilgesellschaft, die maßgeblichen Entscheidungsparameter automatisierter Entscheidungssysteme durch das Crowdsourcing von Daten zu enthüllen, rechtlich zulässig sind (IV.). Abschließend werfe ich einen Blick auf zukünftige externe Kontrolloptionen (V.). Nicht betrachtet werden potentielle Geheimhaltungsmaßnahmen in Gerichtsverfahren, die genug Stoff für einen eigenen Beitrag bieten.[9]

---

[5] *Alba*, Facebook sent flawed data to misinformation researchers, 10.9.2021, https://www.nytimes.com/live/2020/2020-election-misinformation-distortions#facebook-sent-flawed-data-to-misinformation-researchers; *Milmo*, Twitter admits bias in algorithm for rightwing politicians and news outlets, 22.10.2012, https://www.theguardian.com/technology/2021/oct/22/twitter-admits-bias-in-algorithm-for-rightwing-politicians-and-news-outlets.

[6] *Lewis/McCormick*, How an ex-YouTube insider investigated its secret algorithm, 2.2.2018, https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot; *Newton*, Why no one knows which stories are the most popular on Facebook, 22.7.2020, https://www.theverge.com/interface/2020/7/22/21332774/facebook-crowdtangle-kevin-roose-nyt-tweets-interactions-reach-engagement.

[7] *Lapowsky*, Platforms vs. PhDs: How tech giants court and crush the people who study them, 19.3.2021, https://www.protocol.com/nyu-facebook-researchers-scraping; *Briegleb*, Konflikt mit Facebook: Algorithmwatch beendet Instagram-Projekt, 18.8.2021, https://www.heise.de/news/Konflikt-mit-Facebook-Algorithmwatch-beendet-Instagram-Projekt-6165333.html.

[8] *Faife*, Facebook Rolls Out News Feed Change That Blocks Watchdogs from Gathering Data, 21.9.2021, https://themarkup.org/citizen-browser/2021/09/21/facebook-rolls-out-news-feed-change-that-blocks-watchdogs-from-gathering-data; *Waterson*, Facebook restricts campaigners' ability to check ads for political transparency, 27.1.2019, https://www.theguardian.com/technology/2019/jan/27/facebook-restricts-campaigners-ability-to-check-ads-for-political-transparency.

[9] Dazu näher *Gittinger/Redeker/Pres*, WRP 2015, 812; *Hauck*, NJW 2016, 2218 (2221); *Namysłowska*, in: Heermann/Schlingloff (Hrsg.), MüKoUWG (2020), Art. 9 Geheimnisschutz-RL.

II. Der Interessenkonflikt

1. Transparenz

Nicht immer ist unmittelbar einleuchtend, was eine Forderung nach „Transparenz" für automatisierte Entscheidungssysteme beinhaltet. Der Begriff der Transparenz ist schillernd und kann verschiedenste Bedeutungen haben: Die Erkennbarkeit des Einsatzes eines ADM-Systems; die Offenlegung der Entwicklungsziele bzw. der Modellierungsentscheidungen eines ADM-Systems; der Zugang zu Trainings- und Testdaten; die Offenlegung des Quellcodes. Auch eine nachträgliche Blackbox-Analyse durch Dritte oder die Erläuterung von Einzelentscheidungen gegenüber Betroffenen wird unter dem Begriff der Transparenz diskutiert. All diese Maßnahmen können je nach Sachlage sinnvoll und hilfreich sein, um die Funktionsweise eines ADM-Systems zu verstehen und dessen Implikationen für einzelne Betroffene sowie die Gesellschaft insgesamt nachzuvollziehen. Für die Zwecke dieses Beitrags wird deshalb ein weites und umfassendes Verständnis des Begriffs Transparenz zugrunde gelegt, welches – je nach Konstellation – alle der genannten Aspekte umfassen kann.

2. Rechtlicher Vertraulichkeitsschutz

Entwickler und Betreiber von ADM-Systemen halten der Forderung nach Transparenz oftmals ein Interesse an der Geheimhaltung von Informationen entgegen. Rechtlich anerkannt wird ein solches Geheimhaltungsinteresse einerseits durch das Datenschutzrecht, welches personenbezogene Daten schützt, d. h. Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen (vgl. Art. 4 DSGVO). Die Offenlegung personenbezogener Daten durch Übermittlung oder Verbreitung stellt eine Datenverarbeitung dar, die einer Einwilligung des Datensubjekts bedarf oder das Bestehen eines anderen Erlaubnistatbestands nach Art. 6 Abs. 1, 9 Abs. 2 DSGVO voraussetzt. Für Know-How und andere geschäftliche Informationen wird eine rechtliche Anerkennung des Geheimhaltungsinteresses durch das auf der Geschäftsgeheimnis-Richtlinie[10] basierende Gesetz zum Schutz von Geschäftsgeheimnissen (GeschGehG) bewirkt. Nach der Legaldefinition des § 2 Nr. 1 GeschGehG sind Geschäftsgeheimnisse Informationen, die nicht allgemein bekannt oder ohne Weiteres zugänglich sind und deshalb von wirtschaftlichem Wert sind, soweit diese Gegenstand von den Umständen nach angemessenen Geheimhaltungsmaßnahmen durch ihren rechtmäßigen Inhaber sind.

---

[10] Richtlinie (EU) 2016/943 des Europäischen Parlaments und des Rates vom 8. Juni 2016 über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen (Geschäftsgeheimnisse) vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und Offenlegung, Abl. EU Nr. L 157/1 v. 15.6.2016.

3. Anderweitige Geheimhaltungsinteressen

Nicht nur an personenbezogenen Daten und an Geschäftsgeheimnissen können Geheimhaltungsinteressen bestehen. Insbesondere die Abgrenzung zwischen Geschäftsgeheimnissen und anderweitigen Vertraulichkeitsinteressen bereitet Schwierigkeiten. Dies betrifft etwa die Frage, in welchem Umfang faktische Geheimhaltungsmaßnahmen ergriffen werden müssen, um als „angemessen" zu gelten und einen rechtlichen Schutz nach dem GeschGehG entstehen zu lassen.[11] Umstritten ist zudem der rechtliche Schutz von nachteiligen Informationen, wie z. B. Hinweise auf die mangelhafte Funktionsweise eines ADM-Systems[12] oder dessen negative gesellschaftliche Konsequenzen.[13] Nach § 2 Nr. 1 lit. c GeschGehG setzt der Schutz einer Information als Geschäftsgeheimnis ein berechtigtes Interesse an der Geheimhaltung voraus. Doch ist umstritten, ob diese Voraussetzung richtlinienkonform ist.[14] Die Legaldefinition des Geschäftsgeheimnisses in der Geschäftsgeheimnis-RL enthält ein entsprechendes Kriterium nicht explizit. Lediglich Erwägungsgrund 14 erläutert, der Begriff des Geschäftsgeheimnisses solle Know-how, Geschäftsinformationen und technologische Informationen abdecken, bei denen sowohl ein legitimes Interesse an ihrer Geheimhaltung bestehe als auch die legitime Erwartung, dass die Vertraulichkeit gewahrt werde. Im Interesse einer unionsrechtlich einheitlichen Auslegung der Geschäftsgeheimnisrichtlinie sollte Erwägungsgrund 14 allerdings nicht herangezogen werden, um die frühere deutsche Rechtsprechung zur Bagatellschranke des „berechtigten Interesses" zu perpetuieren.

Vielmehr bietet die Legaldefinition der Geschäftsgeheimnis-RL mit dem Merkmal des „wirtschaftlichen Werts" des Geheimnisses einen Ansatzpunkt für eine Interpretation des Schutzumfangs im Lichte von Erwägungsgrund 14.[15] Entscheidend ist die Frage, ob der potentielle[16] wirtschaftliche Wert einen positiven Vermögenswert des Geheimnisinhabers darstellen muss oder ob nur allgemeiner auf das wirtschaftliche Interesse des Geheimnisinhabers abzustellen ist. Würde

---

[11] Näher OLG Hamm, MMR 2021, 506, 508; *Hoeren*, in: Hoeren/Münker, GeschGehG (2021), § 2 Rn. 16 ff.

[12] *Beuth*, Facebooks Forscher zweifeln an Fähigkeiten der Hass-Erkennung, Spiegel Online v. 18.10.2021, https://www.spiegel.de/netzwelt/web/kuenstliche-intelligenz-facebook-zweifelt-anfaehigkeiten-der-hass-erkennung-a-c322d904-da87-4e3b-80b8-ba427186c64d.

[13] Siehe etwa zu dem ADM-System COMPAS, welches das Rückfallrisiko von Straftätern beurteilen soll *Angwin/Larson/Mattu/Kirchner*, Machine Bias – There's software used across the country to predict future criminals. And it's biased against blacks, 23.5.2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[14] Für eine Richtlinienkonformität *Hiéramente*, in: Fuhlrott/Hiéramente, BeckOK GeschGehG (2021), § 2 Rn. 70; von Richtlinienwidrigkeit ausgehend *Alexander*, in: Köhler/Bornkamm/Feddersen, UWG – Kommentar (2021), § 2 GeschGehG Rn. 74.

[15] Zutreffend *Kalbfus*, GRUR 2016, 1009, 1011.

[16] Nach Erwägungsgrund 14 S. 3 Geheimnisschutz-RL bedarf es keines tatsächlichen wirtschaftlichen Werts des Geheimnisses, vielmehr genügt ein entsprechendes Potential.

ein wirtschaftlicher Wert nur dann bejaht, wenn auch eine andere Person als der Geheimnisinhaber aus der positiven Nutzung der geschützten Informationen einen Gewinn ziehen könnte (wie z. B. bei Know-How), so wären Informationen, die sich lediglich negativ auf den Geheimnisinhaber auswirken (und nur reflexiv dessen Konkurrenten begünstigen), nicht schutzfähig.

Meines Erachtens sprechen systematische und teleologische Erwägungen dafür, den potentiellen wirtschaftlichen Wert der Information positiv und unabhängig von der Person des Geheimnisinhabers zu bestimmen. Damit sind nachteilige Informationen, deren Offenbarung dem Geheimnisinhaber einen finanziellen Schaden zufügen können, aus dem Schutzbereich der Richtlinie ausgenommen.[17] Wie sich aus den ersten Erwägungsgründen der Richtlinie ergibt, ist deren vorrangiges Ziel die Innovationsförderung. Unternehmerische Investitionen in die Schaffung und Anwendung „intellektuellen Kapitals" werden als Motor für die Wettbewerbsfähigkeit und den Markterfolg europäischer Unternehmen eingestuft.[18] Der Schutz von Geschäftsgeheimnissen soll deren Inhabern ermöglichen, „einen Nutzen aus ihrer schöpferischen Tätigkeit oder ihren Innovationen zu ziehen".[19] Explizit wird der unionsrechtliche Geheimnisschutz in einen systematischen Zusammenhang zu den Rechtsakten auf dem Gebiet des Geistigen Eigentums gerückt,[20] bei denen die wirtschaftliche Verwertung immaterieller Güter im Vordergrund steht. Vor diesem Hintergrund ist nicht ersichtlich, weshalb auch nachteilige Informationen über schlechte Produktqualitäten, Fehlverhalten oder gar rechtswidrige Aktivitäten unter den zusätzlichen rechtlichen Schutz des Geschäftsgeheimnisrechts gestellt werden sollten.[21] Mit Blick auf solche Informationen steht den Geheimniswahrern einerseits die Möglichkeit offen, faktische Schutzmaßnahmen zu ergreifen; andererseits können arbeitsvertragliche und sonstige Verschwiegenheitsvereinbarungen getroffen werden.

Dagegen wird in der Literatur teilweise eingewandt, die in der Richtlinie vorgesehene Ausnahme zugunsten von Whistleblowern wäre entbehrlich, wenn Informationen über rechtswidriges Verhalten ohnehin nicht dem Geheimnisschutz unterfielen.[22] Doch ist dies kein schlagendes Argument: Regelmäßig sind Infor-

---

[17] Im Ergebnis auch *Kalbfus*, GRUR 2016, 1009, 1011. A. A. *Alexander*, in: Köhler/Bornkamm/Feddersen, UWG – Kommentar (2021), § 2 GeschGehG Rn. 79; *Ullrich*, NZWiSt 2019, 65, 67.

[18] Erwägungsgrund 1 Geheimnisschutz-RL.

[19] Erwägungsgrund 2 Geheimnisschutz-RL.

[20] Erwägungsgründe 1 und 2 Geheimnisschutz-RL.

[21] Zu den schwerwiegenden Wertungswidersprüchen, die aus einem Geheimnisschutz für rechtswidrige Umstände oder rechtswidrige Vorgänge resultieren würden, siehe *Alexander*, in: Köhler/Bornkamm/Feddersen, UWG – Kommentar (2021), § 2 GeschGehG Rn. 79 mit weiteren Nachweisen in Fn. 78; für den Einschluss von Informationen zu rechtswidrigen Handlungen unter den Begriff des Geschäftsgeheimnisses *Hiéramente*, in: Fuhlrott/Hiéramente (Hrsg.), BeckOK GeschGehG (Juni 2021), § 2 Rn. 72 f. m. w. N.; *Ullrich*, NZWiSt 2019, 65, 67; *Schröder*, ZRP 2020, 212, 214.

[22] *Ullrich*, NZWiSt 2019, 65, 67; *Hiéramente*, in: Fuhlrott/Hiéramente, BeckOK GeschGehG (Juni 2021), § 2 Rn. 73, 73.1.; in diese Richtung auch *Ohly*, GRUR 2019, 441, 444 f.

mationen über rechtswidriges Verhalten oder anderweitige nachteilige Tatsachen mit Inhalten verquickt, an deren Geheimhaltung ein legitimes Interesse besteht.[23] Dies gilt insbesondere dann, wenn der oder die Whistleblower Dokumente vorlegen, um ihre Behauptungen zu untermauern. So enthalten beispielsweise die dem Wall Street Journal zugespielten Dokumente zur schlechten Funktionsweise der von Facebook eingesetzten Hassrede-Filter auch Informationen über die von Facebook aufgestellte Kostenkalkulation menschlicher Inhaltemoderation.[24]

Bis zu einer Auslegungsentscheidung des EuGH ist letztlich offen, wie weit der durch die Richtlinie vorgegebene Begriff des Geschäftsgeheimnisses zu ziehen ist.[25] Abhängig hiervon gibt es Geheimhaltungsinteressen auch an einem Kreis von Informationen, die nicht als Geschäftsgeheimnis einzuordnen sind. Dies betrifft die genannten Konstellationen, in denen das Bekanntwerden der Informationen zu Reputationsverlust führen könnte. Daneben wird es auch Situationen geben, in denen den Betreiber eines ADM-Systems vertragliche Verschwiegenheitspflichten treffen. Schließlich kann ein Interesse daran bestehen, die Manipulation eines ADM-Systems durch Dritte zu verhindern, ohne dass ein Geheimnisschutz nach dem GeschGehG existiert. Solche Geheimhaltungsinteressen mögen nicht mit einem quasi-immaterialgüterrechtlichen Schutz ausgestattet sein, können aber gleichwohl rechtliche Berücksichtigung finden, z. B. bei der Ausgestaltung gesetzlicher Offenbarungspflichten oder im Kontext der AGB-Kontrolle.

4. Relevante Entwicklungszyklen

Geheimhaltungsinteressen können in jedem Stadium des Entwicklungszyklus eines ADM-Systems auftreten: Bereits in der Planungsphase eines ADM-Systems ebenso wie bei Festlegung der Modellierungsentscheidungen entstehen regelmäßig Geschäftsgeheimnisse. Im Zuge der Datenerhebung und -aufbereitung sowie der Trainings- und Testphase werden oftmals personenbezogene oder unternehmensbezogene Daten verarbeitet. Gelangt das ADM-System zur Anwendung in der Praxis, wird häufig die Vermeidung von Manipulationen ein wichtiges Ziel sein, und im Zuge der Evaluation kann insbesondere ein Interesse an der Geheimhaltung nachteiliger Informationen bestehen, beispielsweise, wenn die Leistungsstärke des Systems eher gering ist.

---

[23] Siehe auch *Alexander*, in: Köhler/Bornkamm/Feddersen, UWG – Kommentar (2021), § 2 GeschGehG Rn. 79, 82.

[24] *Seetharaman/Horwitz/Scheck*, Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts, Wall Street Journal v. 17.10.2021, https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184?mod=article_inline.

[25] Zwar bewirkt die Richtlinie ausweislich ihres Art. 1 Abs. 1 S. 1 nur eine Mindestharmonisierung, doch hat der deutsche Gesetzgeber den Kreis der geschützten Informationen nicht weiter ziehen wollen als durch Art. 2 Nr. 1 der Richtlinie vorgegeben.

III. Informations- und Offenlegungspflichten

1. Status Quo

Ein Einblick in die Funktionsweise von ADM-Systemen lässt sich zunächst durch den Einsatz von Informations- und Offenbarungspflichten erzielen. Die Regelungen des Daten- bzw. Geheimnisschutzes stehen der Einführung solcher Pflichten nicht entgegen.[26] Eine Offenlegung der Funktionsweise eines ADM-Systems wird in der Regel möglich sein, ohne personenbezogene Daten preiszugeben. Davon abgesehen enthalten Art. 6 Abs. 1 lit. c, Art. 9 Abs. 2 lit. g DSGVO Erlaubnistatbestände für die Verarbeitung von personenbezogenen Daten zwecks der Erfüllung rechtlicher Verpflichtungen. Das GeschGehG wiederum schützt nur vor der rechtswidrigen Erlangung, Nutzung oder Offenlegung einer Information. Die Einführung gesetzlicher Offenlegungspflichten ist nach Art. 3 Abs. 2 Geschäftsgeheimnis-RL auch durch mitgliedstaatliches Recht erlaubt.

In Bezug auf ADM-Systeme existieren allerdings gegenwärtig nur wenige entsprechende Offenbarungspflichten. So enthält die DSGVO für den Fall einer automatisierten Entscheidungsfindung mit rechtlicher oder äquivalenter Wirkung eine Informationspflicht gegenüber bzw. ein Auskunftsrecht der hiervon betroffenen natürlichen Person. Zur Verfügung gestellt werden sollen „aussagekräftige Informationen über die involvierte Logik" der automatisierten Entscheidungsfindung (Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g, Art. 15 Abs. 1 lit. h DSGVO). Ansprüche auf Auskunft über die Daten, mit welchen das Entscheidungssystem trainiert wurde, auf die Offenlegung des Quellcodes oder auf Begründung einer getroffenen Entscheidung lassen sich hieraus nicht ableiten.[27] Auch eine Darlegung der Funktionsweise des ADM-Systems im Detail ist nicht erforderlich. Allenfalls besteht ein Anspruch auf Unterrichtung über die Grundsätze der Modellierungsentscheidungen; bei weiter Auslegung auch über die Quellen der verwendeten Daten.[28]

Ähnlich oberflächlich bleiben die Darstellungspflichten aus Art. 5 P2B-VO[29] und § 5b Abs. 2 UWG, wonach Online-Vermittlungsdienste und Suchmaschinenbetreiber, die Rankings erstellen, die das Ranking bestimmenden Hauptparameter und die relative Gewichtung dieser Hauptparameter gegenüber anderen Parame-

---

[26] Siehe für den Forschungsdatenzugang *Specht-Riemenschneider*, Studie zur Regulierung eines privilegierten Zugangs zu Daten für Wissenschaft und Forschung durch die regulatorische Verankerung von Forschungsklauseln in den Sektoren Gesundheit, Online-Wirtschaft, Energie und Mobilität, August 2021, https://www.jura.uni-bonn.de/fileadmin/Fachbereich_Rechtswissenschaft/Einrichtungen/Lehrstuehle/Specht/Dateien/2021-08-25-LSR.pdf, S. 5.

[27] Zu Entscheidungsbegründungen siehe *Wachter/Mittelstadt/Russell*, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law & Technology, 31 (2), 2018, S. 872.

[28] A. A. wohl *Hacker*, Common Market Law Review 2018, 1143, 1173 f.

[29] Verordnung (EU) 2019/1150 des Europäischen Parlaments und des Rates vom 20. Juni 2019 zur Förderung von Fairness und Transparenz für gewerbliche Nutzer von Online-Vermittlungsdiensten, ABL. L 186/57 v. 11.7.2019.

tern darzustellen haben. Schließlich sind nach § 2 Abs. 2 NetzDG „Art, Grundzüge der Funktionsweise und Reichweite" von ADM-Systemen zu erläutern, die von sozialen Netzwerken zur Filterung der von ihnen gespeicherten Inhalte eingesetzt werden, wobei auch „allgemeine Angaben zu den verwendeten Trainingsdaten sowie zur Überprüfung" der Systeme zu machen sind. Diese Offenbarungspflichten beschränken sich jeweils auf allgemeinere Aussagen und ermöglichen keine äußere Kontrolle der Funktionsweise von ADM-Systemen.

2. Vorschläge de lege ferenda

Ähnlich der Regelung in der P2B-Verordnung sieht Art. 29 Abs. 1 des Entwurfs eines Digital Services Acts für von sehr großen Online-Plattformen eingesetzte algorithmische Empfehlungssysteme eine Offenlegung der „wichtigsten Parameter" vor. Auch der von der Europäischen Kommission im April 2021 veröffentlichte Verordnungsvorschlag für ein Gesetz über künstliche Intelligenz[30] geht insoweit kaum weiter. Zwar treffen die Entwickler von ADM-Systemen nach Art. 11 KI-VO-E umfangreiche technische Dokumentationspflichten. Die in Art. 13 KI-VO-E vorgesehenen Transparenzregelungen zielen allerdings vorrangig auf die Erkennbarkeit des Einsatzes bestimmter ADM-Systeme[31] sowie auf Informationen, die den Nutzer:innen den Gebrauch des Systems ermöglichen und ihnen erlauben, die mit dessen Einsatz verbundenen Risiken abzuschätzen.[32] „Gegebenenfalls" sind zudem nach Art. 13 Abs. 3 lit. b (v) KI-VO-E die „Spezifikationen für die Eingabedaten oder sonstige[r] relevante[r] Informationen über die verwendeten Trainings-, Validierungs- und Testdatensätze unter Berücksichtigung der Zweckbestimmung des KI-Systems" mitzuteilen. Ohnehin beschränken sich diese Regelungen auf im Verordnungsvorschlag näher spezifizierte Hochrisiko-Algorithmen[33] und erfassen deshalb viele ADM-Systeme nicht (z. B. Rankings auf Vergleichsplattformen, Empfehlungssysteme sozialer Netzwerke).

3. Bewertung

Gesetzliche Offenbarungspflichten können einen gewissen Beitrag zur Transparenz von ADM-Systemen leisten. Doch kann durchaus ein berechtigtes Interesse an der Geheimhaltung des dem System zugrunde liegenden Know-Hows bestehen, die eine Offenbarungspflicht nicht adäquat erscheinen lässt. Zudem gilt es stets abzuwägen, ob durch die Offenlegung von Informationen eine Option geschaffen wird, das System zu manipulieren. Ein solches Interesse an der Manipu-

---

[30] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz, 21.4.2021, COM(2021) 206.

[31] Art. 52 KI-VO-E.

[32] Art. 13 Abs. 3 i. V. m. Erwägungsgrund 47 KI-VO-Entwurf.

[33] Art. 6 KI-VO-E.

lation der Funktionsweise durch Dritte besteht beispielsweise bei Rankings, Empfehlungs- und Scoring-Systemen. Selbst dort, wo Offenbarungspflichten bestehen, fällt es von außen schwer zu beurteilen, ob die von dem Hersteller oder Betreiber erteilten Informationen über Entscheidungsparameter und deren Gewichtung der Realität entsprechen. Ohne korrespondierende Befugnisse der Behörden, die Korrektheit der Informationen zu überprüfen, laufen Offenbarungspflichten letztlich ins Leere. Dies gilt selbstverständlich auch für freiwillig zur Verfügung gestellte Informationen, wie z. B. wissenschaftliche Vorträge und Aufsätze aus dem Kreis der Belegschaft des systementwickelnden Unternehmens oder solche Informationen, die mittels Analyseinstrumenten des ADM-Systembetreibers ausgelesen werden können.[34]

IV. Zulässigkeit faktischer Transparenzmechanismen

1. Unautorisierte Blackbox-Analyse

Da die Informations- und Offenbarungspflichten der Betreiber von ADM-Systemen einem Papiertiger gleichkommen, wird von verschiedenster Seite versucht, die Funktionsweise von ADM-Systemen durch externes Testen zu durchleuchten. Ich bezeichne diese Maßnahmen im Folgenden als unautorisierte Blackbox-Analyse. Dabei verweist der Begriff „Blackbox" nicht darauf, dass die untersuchten Systeme per se nicht nachvollziehbar wären (wie dies bei deep learning-Systemen der Fall sein kann[35]). Vielmehr soll damit ausgedrückt werden, dass sich die Analyst:innen den ADM-Systemen ohne nähere Informationen und ohne die Zustimmung des systementwickelnden Unternehmens nähern. Ein solches Vorgehen setzt entweder die „Datenspende" betroffener Personen voraus, wie beispielsweise bei den Projekten OpenSCHUFA und Citizen Browser, die die Ermittlung der SCHUFA-Score-Berechnung bzw. des Facebook Newsfeed-Algorithmus zum Ziel hatten bzw. haben. Bei OpenSCHUFA hatte der Projektträger AlgorithmWatch Datensubjekte dazu aufgerufen, bei der SCHUFA eine Selbstauskunft zu beantragen und die erlangten Informationen an AlgorithmWatch weiterzuleiten.[36] Bei Datenspenden zwecks der Analyse von sozialen Netzwerken werden den Datenspender:innen in der Regel spezifische Webbrowser oder Browser-Erweiterungen zur Verfügung gestellt, um soziale Netzwerke zu nutzen.[37] Alternativ zu

---

[34] Z. B. die Facebook Ad Library oder das Facebook-Analyseinstrument CrowdTangle. Zu letzterem näher *Roose*, Inside Facebook's Data Wars, 14.7.2021, https://www.nytimes.com/2021/07/14/technology/facebook-data.html.

[35] Näher https://gi.de/informatiklexikon/explainable-ai-ex-ai.

[36] Nähere Informationen unter https://openschufa.de.

[37] Siehe näher https://themarkup.org/citizen-browser; https://algorithmwatch.org/de/monitoring-instagram/; https://adobserver.org/.

einer solchen Datenspende können die Analyst:innen unechte Nutzer:innenprofile anlegen, beispielsweise zwecks Analyse der von sozialen Netzwerken eingesetzten Empfehlungsalgorithmen.[38]

Namentlich Facebook geht durch die Sperrung von Nutzerkonten und die Androhung rechtlicher Schritte gegen solche unautorisierten Analyseversuche vor.[39] Aufgrund der Einwilligung der Datenspender stellt das Datenschutzrecht kein Hindernis für eine Blackbox-Analyse dar.[40] Selbiges gilt für das Urheberrecht,[41] sofern das eingesetzte Computerprogramm nicht vervielfältigt wird.[42] Im Folgenden will ich deshalb beleuchten, ob die unautorisierte Blackbox-Analyse die Verletzung eines Geschäftsgeheimnisses darstellt oder jedenfalls durch Allgemeine Geschäftsbedingungen unterbunden werden kann.

2. Beurteilung nach dem Geschäftsgeheimnisgesetz

a) Verletzungstatbestand

Halten Entwickler und Betreiber eines ADM-Systems dessen genaue Funktionsweise mittels angemessener Vertraulichkeitsmaßnahmen geheim und kommt der Funktionsweise – wie regelmäßig – ein jedenfalls potentieller wirtschaftlicher Wert zu, so handelt es sich hierbei um ein Geschäftsgeheimnis i. S. d. § 2 Nr. 1 GeschGehG. Damit entsteht die Frage, ob eine unautorisierte Blackbox-Analyse zwecks Identifikation der Entscheidungsparameter des Systems eine Verletzung

---

[38] *Lapowsky*, Platforms vs. PhDs: How tech giants court and crush the people who study them, 19.3.2021, https://www.protocol.com/nyu-facebook-researchers-scraping.

[39] *Vincent*, Facebook bans academics who researched ad transparency and misinformation on Facebook, 4.8.2021, https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in; *Briegleb*, Konflikt mit Facebook: Algorithmwatch beendet Instagram-Projekt, 18.8.2021, https://www.heise.de/news/Konflikt-mit-Facebook-Algorithmwatch-beendet-Instagram-Projekt-6165333.html.

[40] Art. 6 Abs. 1 lit. a, 9 Abs. 2 lit. a DSGVO.

[41] Die Blackbox-Analyse an sich stellt weder eine Verletzung des Urheberrechts noch eine Verletzung verwandter Schutzrechte dar, siehe Entwurf eines Gesetzes zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft, BT-Drucks. 18/12329 v. 15.5.2017, S. 40; *Hagemeier*, in: Ahlberg/Götting/Lauber-Rönsberg (Hrsg.), BeckOK Urheberrecht, § 60d Rn. 5; *Hentsch*, in Dreyer/Kotthoff/Meckel/Hentsch (Hrsg.), Heidelberger Kommentar UrheberR (2018), § 60d Rn. 3; *Stieper*, in: Schricker/Loewenheim, UrhG, 6. Aufl. 2020, § 60d Rn. 3; *Dreier*, in: Dreier/Schulze, UrhG, 6. Aufl. 2018, § 60d Rn. 4. Die Datensammlung durch Scraping kann allerdings in das Recht des Datenbankherstellers eingreifen, wenn die Beschaffung, Überprüfung oder Darstellung der Daten durch den Datenbankhersteller eine nach Art oder Umfang wesentliche Investition erfordert hat und ein nach Art oder Umfang wesentlicher Teil der Datenbank vervielfältigt wird (§§ 87a Abs. 1 S. 1, 87b Abs. 1 S. 1 UrhG). Selbst in diesem Fall erlaubt jedoch die Schranke für Text und Data Mining gemäß § 60d UrhG die Vervielfältigung der Datenbank bzw. urheberrechtlich geschützter Werke für die Zwecke der wissenschaftliche Forschung, vgl. *Raue*, Die geplanten Text und Data Mining-Schranken (§§ 44b und 66d UrhG-E) ZUM 2020, 172 ff., auch zum Text und Data Mining für nichtwissenschaftliche Zwecke gemäß § 44b UrhG.

[42] §§ 69a, 69c UrhG.

des Geschäftsgeheimnisses nach § 4 Abs. 1 GeschGehG darstellt. Dies setzt voraus, dass die Parameter identifiziert werden durch a) unbefugten Zugang zu Dateien, die der rechtmäßigen Kontrolle des Inhabers des Geschäftsgeheimnisses unterliegen und aus denen sich das Geschäftsgeheimnis ableiten lässt oder b) ein sonstiges Verhalten, das unter den jeweiligen Umständen nicht dem Grundsatz von Treu und Glauben unter Berücksichtigung der anständigen Marktgepflogenheit entspricht.

b) Positivliste: Reverse Engineering

Die Interpretation des Verletzungstatbestands des § 4 GeschGehG ist im Lichte der Positivliste erlaubter Handlungen nach § 3 GeschGehG vorzunehmen. Erlaubt ist danach unter anderem die Erlangung eines Geschäftsgeheimnisses durch das so genannte *reverse engineering*. Dieses wird definiert als ein Beobachten, Untersuchen, Rückbauen oder Testen eines Produkts oder Gegenstands, das oder der öffentlich verfügbar gemacht wurde oder sich im rechtmäßigen Besitz des Analysierenden befindet (sofern letzterer keiner Pflicht zur Beschränkung der Erlangung des Geschäftsgeheimnisses unterliegt).[43] Die Definition des *reverse engineering* wurde ersichtlich mit Blick auf körperliche Gegenstände konzipiert, so dass bereits zweifelhaft ist, ob sich auch unkörperliche Gegenstände unter den Gegenstandsbegriff der Norm subsumieren lassen.[44] Beispielsweise kann der in § 3 Abs. 1 Nr. 2 lit. b angesprochene „Besitz" des untersuchten Gegenstands grundsätzlich nur an körperlichen Gegenständen erlangt werden. Doch wäre die Differenzierung zwischen „Produkt" und „Gegenstand" in § 3 Abs. 1 Nr. 2 GeschGehG entbehrlich, wenn außer körperlichen Gegenständen nicht auch unkörperliche Gegenstände erfasst wären. Auch aus der Sonderregel zur Dekompilation von Software nach § 69e UrhG lässt sich kein Argument gegen eine Anwendung des § 3 Abs. 1 Nr. 2 GeschGehG auf ADM-Systeme ableiten. Denn §§ 69d Abs. 3, 69e UrhG regeln lediglich die urheberrechtliche Zulässigkeit im Falle der Vornahme urheberrechtlich relevanter Nutzungsvorgänge, insbesondere bei Vervielfältigung eines Computerprogramms, die für die Analyse eines ADM-Systems nicht zwingend erforderlich ist (falls doch, wären diese Bestimmungen zusätzlich zu prüfen). Eine weite Interpretation des *reverse engineering* nach § 3 Abs. 1 Nr. 2 GeschGehG, die auch die Untersuchung unkörperlicher Gegenstände erfasst, ist somit zu befürworten.

---

[43] § 3 Abs. 1 Nr. 2 GeschGehG.

[44] Für eine weite Auslegung des Begriffs *Alexander*, in: Köhler/Bornkamm/Feddersen, UWG – Kommentar (2021), § 3 GeschGehG Rn. 29; *Drescher*, in: Hoeren/Sieber/Holznagel, MMR-Hdb (56. EL Mai 2021),Teil 7.9 Rn. 20; *Strobel*, in: Hoeren/Münker, GeschGehG (2021), § 3 Abs. 1 Nr. 2 Rn. 45; a. A. *Spieker*, in: Fuhlrott/Hiéramente, BeckOK GeschGehG (2021), § 3 Rn. 11, wonach sich der Gegenstandsbegriff nur auf körperliche Sachen i. S. d. § 90 BGB bezieht.

Im Einzelfall ist folglich zu ermitteln, ob das untersuchte ADM-System i. S. d. § 3 Abs. 1 Nr. 2 lit. a GeschGehG „öffentlich verfügbar gemacht wurde". Öffentlichkeit ist dabei im Sinne einer unbestimmten Anzahl von Personen zu verstehen, die nicht durch engere persönliche oder geschäftliche Beziehungen zueinander verbunden sind.[45] Von einer Verfügbarmachung ist regelmäßig auszugehen, wenn das Erzeugnis auf den Markt und in den Vertrieb gelangt ist.[46] Bei ADM-Systemen sollte hierfür unerheblich sein, ob die Software selbst zur Verfügung gestellt wird oder nur deren Nutzung (entgeltlich oder unentgeltlich) der Öffentlichkeit ermöglicht wird. Denn Ziel der Untersuchung ist nicht die Ermittlung des Quellcodes, sondern die Funktionsweise des ADM-Systems. Ist das ADM-System nur einem begrenzten Personenkreis zur Verfügung gestellt worden, bietet sich mit Blick auf die alternative erlaubte Handlung des rechtmäßigen, unbeschränkten Besitzes nach § 3 Abs. 1 Nr. 2 lit. b eine weite Auslegung des Besitzbegriffs an – im Sinne einer tatsächlichen Zugriffsmöglichkeit auf das Entscheidungssystem.[47]

An einer öffentlichen Verfügbarmachung eines ADM-Systems fehlt es jedoch, sofern nur die Ergebnisse der ADM-Entscheidung oder -Empfehlung dem Datensubjekt bzw. interessierten Dritten mitgeteilt werden, wie dies etwa beim individuellen SCHUFA-Score oder einem medizinischen Befund der Fall ist. Auch von einem rechtmäßigen Besitz des untersuchten Gegenstands kann in diesem Fall nicht die Rede sein. Allenfalls werden die von den Datenspendern weitergeleiteten Daten rechtmäßig verarbeitet. Zwar kann der Inhaber eines Geschäftsgeheimnisses über die Funktionsweise eines ADM-Systems natürlichen Personen nicht untersagen, ihre personenbezogenen Daten weiterzugeben. Dies bedeutet allerdings nicht, dass es einem Dritten erlaubt wäre, mittels dieser Daten *reverse engineering* eines ADM-Systems zu betreiben.

c) Nochmals: Verletzungstatbestand

Eine Subsumtion der unautorisierten Blackbox-Analyse unter die erlaubte Handlung des *reverse engineering* nach § 3 Abs. 1 Nr. 2 ist allerdings keine notwendige Bedingung, um die Analyse als erlaubt anzusehen. Denn den Tatbeständen des § 3 Abs. 1 GeschGehG kommt nur eine klarstellende Funktion zu.[48] Entscheidend ist, ob eines der Handlungsverbote des § 4 Abs. 1 GeschGehG einschlägig ist. Nach § 4 Abs. 1 Nr. 1 GeschGehG darf ein Geschäftsgeheimnis unter anderem nicht

---

[45] Ähnlich *Spieker*, in: Fuhlrott/Hiéramente, BeckOK GeschGehG (2021), § 3 Rn. 14; *Strobel*, in: Hoeren/Münker, GeschGehG (2021), § 3 Abs. 1 Nr. 2 Rn. 52.

[46] Begr. RegE, BT-Drs. 19/4724, 26; *Alexander*, in: Köhler/Bornkamm/Feddersen, UWG – Kommentar (2021), § 3 GeschGehG Rn. 32; vgl. *Spieker*, in: Fuhlrott/Hiéramente, BeckOK GeschGehG (2021), § 3 Rn. 13 f.

[47] Näher zum unionsrechtlichen Begriff des Besitzes *Alexander*, in: Köhler/Bornkamm/Federsen, UWG – Kommentar (2021), § 3 GeschGehG, Rn. 35.

[48] *Alexander*, in: Köhler/Bornkamm/Federsen, UWG – Kommentar (2021), § 3 GeschGehG, Rn. 9.

erlangt werden durch den unbefugten Zugang zu bzw. die unbefugte Aneignung von Dokumenten und elektronischen Dateien, die der rechtmäßigen Kontrolle des Inhabers des Geschäftsgeheimnisses unterliegen. Dies ist freilich bei Datenspenden nicht der Fall: Hier liegt die rechtmäßige Kontrolle über die Daten jedenfalls auch bei den Datenspender:innen, d. h. bei denjenigen Personen, welche im Zuge der Nutzung des ADM-Systems diese Daten generieren bzw. die betreffenden Informationen durch Ausübung ihrer datenschutzrechtlichen Auskunftsrechte erlangt haben.

Eine Geheimnisverletzung im Wege der unautorisierten Blackbox-Analyse kann sich damit allenfalls noch aus § 4 Abs. 1 Nr. 2 GeschGehG ergeben. Danach ist die Erlangung eines Geschäftsgeheimnisses rechtswidrig, wenn das jeweilige Verhalten unter den konkreten Umständen nicht dem Grundsatz von Treu und Glauben unter der Berücksichtigung anständiger Marktgepflogenheiten entspricht. Werden die zur Entschlüsselung des Entscheidungssystems erforderlichen Daten im Wege eines erlaubten API-Zugriffs, aufgrund von freiwilligen Datenspenden oder durch das Anlegen von Nutzerkonten erlangt, fehlt es jedoch an Anhaltspunkten, die einen Widerspruch zu marktadäquatem Verhalten begründen. Selbst wenn im Einzelfall ein Verstoß gegen die Allgemeinen Geschäftsbedingungen des Betreibers des ADM-Systems erfolgt, ist dies nicht als treuwidrig einzustufen, sofern nur eine Analyse vorgenommen wird, die nicht mit einer Schädigung des Systems einhergeht.[49] Eine Geheimnisverletzung scheidet mithin unabhängig davon aus, ob man die Tatbestandsausnahme des *reverse engineering* in den hier interessierenden Sachverhaltskonstellationen für anwendbar erachtet.

3. Unterbindung der unautorisierten Blackbox-Analyse in AGB

Auch wenn das Geschäftsgeheimnisgesetz einer unautorisierten Blackbox-Analyse nicht entgegensteht, können sich entsprechende Beschränkungen gegebenenfalls aus den Allgemeinen Geschäftsbedingungen des Unternehmens ergeben, welches das ADM-System veräußert oder zur Nutzung bereitstellt. Verstöße gegen die AGB können zur Sperre der für die Analyse erforderlichen Nutzerkonten führen.[50] Im Folgenden werden drei potentielle Klauselinhalte betrachtet: (1) Beschränkungen des *reverse engineering*, (2) Verbote, fiktive Nutzerkonten anzulegen sowie (3) Verbote von Datenspenden, ggf. im Wege der automatisierten Datenextrahierung (sog. *Scraping*).

---

[49] Siehe zur Parallelwertung im Lauterkeitsrecht BGH v. 12.1.2017 – I ZR 253/14, GRUR 2017, 397 – World of Warcraft II, Rn. 68.

[50] So etwa im Fall der Kontensperre von Forscher:innen des NYU Ad Observatory, siehe *Vincent*, Facebook bans academics who researched ad transparency and misinformation on Facebook, 4.8.2021, https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in.

AGB-rechtliche Beschränkungen des *reverse engineering* sind gemäß § 307 Abs. 1, Abs. 2 Nr. 1 BGB an den wesentlichen Grundgedanken der gesetzlichen Regelung in § 3 Abs. 1 Nr. 2 GeschGehG zu messen. Wie bereits oben erwähnt, differenziert diese Regelung zwischen der Analyse von Gegenständen, die öffentlich verfügbar gemacht wurden sowie der Analyse von Gegenständen, die sich im rechtmäßigen Besitz einer Person befinden, die keiner Pflicht zur Beschränkung der Erlangung des Geschäftsgeheimnisses unterliegt. Aus dieser Differenzierung ergibt sich unmittelbar, dass eine vertragliche Beschränkung des *reverse engineering* nur in jenen Fällen in Betracht kommt, in denen der Gegenstand nicht öffentlich verfügbar gemacht wurde.[51] Eine stärkere Einschränkung der Analyse durch Allgemeine Geschäftsbedingungen ist mit den wesentlichen Grundgedanken der gesetzlichen Regelung nicht vereinbar und stellt eine unangemessene Benachteiligung des Vertragspartners dar, die nach § 307 Abs. 1 BGB unwirksam ist.

Kann zwar die Blackbox-Analyse per se nicht über die Grenzen des GeschGehG hinaus beschränkt werden, so ist es aber gegebenenfalls zulässig, den Zugriff auf die für die Analyse erforderlichen Daten in Nutzungsbedingungen auszuschließen. Gebräuchlich sind beispielsweise in sozialen Netzwerken Klarnamenpflichten sowie Klauseln, die das Erstellen mehrerer, fiktiver oder automatisierter Nutzerkonten verbieten. Der AGB-Kontrolle halten entsprechende Klauseln nur stand, wenn sie die andere Vertragspartei nicht entgegen den Geboten von Treu und Glauben unangemessen benachteiligen (§ 307 Abs. 1 BGB). Entscheidend ist, ob der AGB-Verwender durch eine einseitige Vertragsgestaltung missbräuchlich eigene Interessen zulasten seiner Vertragspartner durchsetzen will, ohne einen angemessenen Ausgleich zwischen den Parteien anzustreben.

Die AGB-rechtliche Zulässigkeit einer Klarnamenpflicht in Telemedien ist in Rechtsprechung und Schrifttum umstritten,[52] weil § 19 Abs. 2 TTDSG den Nutzer:innen grundsätzlich ein Recht auf eine pseudonyme Nutzung von Telemedien gewährt.[53] Der BGH geht von einer Klarnamenpflicht gegenüber dem Diensteanbieter und der Möglichkeit zur pseudonymen Nutzung des Dienstes im Außenverhältnis aus.[54] Für die Zulässigkeit der Verwendung fiktiver Nutzerkonten zu Forschungs-

---

[51] *Leister*, Liberalisierung von Reverse Enginieering durch Geschäftsgeheimnisgesetz: Wie können sich Unternehmen noch schützen?, GRUR-PRax 2019, 175, 176; *Reinfeld*, Das neue Gesetz zum Schutz von Geschäftsgeheimnissen, § 2 Rn. 32; *Strobel*, in: Hoeren/Münker, GeschGehG (2021), § 3 Abs. 1 Nr. 2 Rn. 60 f. hält lediglich eine individualvertragliche Vereinbarung für zulässig.

[52] Für die Zulässigkeit einer entsprechenden Klausel: OLG München v. 8.12.2020 – 18 U 2822/19, MMR 2021, 245 Rn. 40 ff.; *Nebel*, Die Zulässigkeit der Erhebung des Klarnamens nach den Vorgaben der Datenschutz-Grundverordnung K&R 2019, 148 ff.; einschränkend befürwortend: *Gräfe/Hamm*, Anonymität im Internet, in: Berger/Deremetz/Henning/Michell (Hrsg.), Autonomie und Verantwortung in digitalen Kulturen, S. 266 f.; gegen die Zulässigkeit einer Klarnamenpflicht *Caspar*, Klarnamenpflicht versus Recht auf pseudonyme Nutzung, ZRP 2015, 233 ff.

[53] Zur Vereinbarung der vormals in § 13 Abs. 6 S. 1 TMG enthaltenen Regelung mit den Bestimmungen der DSGVO siehe OLG München v. 8.12.2020 – 18 U 2822/19, MMR 2021, 245 Rn. 49 ff. – Klarnamenpflicht.

[54] BGH v. 27.1.2022 – III ZR 3/21, MMR 2022, 375 Rn. 39 ff.

zwecken ist diese Entscheidung freilich nicht ausschlaggebend, da das Erstellen von Profilen, die eine falsche Identität vorspiegeln, in Nutzungsbedingungen auch unabhängig von einer Klarnamenpflicht untersagt werden könnte.[55] Insofern ist zu berücksichtigen, dass die Betreiber von ADM-Systemen ein berechtigtes Interesse daran haben können, eine nicht authentische Nutzung ihrer Dienste zu unterbinden, um eine Irreführung und Manipulation anderer Nutzer:innen zu vermeiden und wahrheitsgemäße Angaben zu ihrem Nutzerkreis gegenüber Investor:innen und Werbepartner:innen sicherzustellen. In den AGB ist folglich das berechtigte Interesse an der Erforschung der Wirkweise von algorithmischen Empfehlungssystemen gegenüber den Risiken einer inauthentischen Nutzung je nach ADM-System abzuwägen. Ein Verbot der Anlage fiktiver Nutzerkonten stellt nicht zwingend eine missbräuchliche Durchsetzung der eigenen Interessen des Systembetreibers dar.

Grundsätzlich anders stellt sich die Interessenabwägung in den Konstellationen einer simplen Datenspende dar. Wird die personenbezogene Empfehlung eines ADM-Systems einer natürlichen Person mitgeteilt, wie beispielsweise ein SCHUFA-Score oder eine medizinische Beurteilung, so kann die Weiterleitung dieser Daten in Anbetracht der Datenautonomie der betroffenen Person nicht unterbunden werden. Eine Verarbeitung zu Zwecken der Blackbox-Analyse ist mit Einwilligung der betroffenen Person nach Art. 6 Abs. 1 Nr. 1 DSGVO zulässig. Ein berechtigtes Interesse des Systembetreibers, eine solche Datenspende zu unterbinden, ist nicht erkennbar.

Wiederum anders gestaltet sich die Interessenlage im Fall der automatisierten Datenextrahierung durch eine nutzerseitig eingesetzte Software (Browser bzw. Browser-Erweiterung). Der Rechtsgedanke des Anspruchs auf Datenportabilität gemäß Art. 20 DSGVO[56] streitet auf den ersten Blick dafür, dass Nutzer:innen die im Kontext des Entscheidungssystems erzeugten Daten aufzeichnen und an andere verantwortliche Stellen weiterleiten dürfen. In der Datenökonomie sollte es Nutzer:innen erlaubt sein, die Verarbeitung der eigenen personenbezogenen Daten zu monetarisieren oder altruistisch der Forschung zur Verfügung zu stellen. Freilich stehen einer solchen automatisierten Datenextrahierung auch erhebliche datenschutzrechtliche Bedenken entgegen. So hat der Cambridge Analytica-Skandal gezeigt, dass die Akquise von Datenspenden zu vorgeblich wissenschaftlichen Zwecken missbraucht werden kann, um die Daten zu Marketingzwecken zu ver-

---

[55] LG Frankfurt v. 3.9.2020 – 2-03 O 282/19 (juris). Zur Frage, ob das Erstellen fiktiver Nutzerkonten einen Verstoß gegen den US-amerikanischen *Computer Fraud and Abuse Act* darstellt, siehe verneinend *Sandvig v. Barr*, D.D.C., No. 16-1368 (JDB), https://www.aclu.org/sites/default/files/field_document/sandvig_opinion.pdf.

[56] Eine direkte Anwendung des Art. 20 DSGVO kommt regelmäßig nicht in Betracht, da die Bestimmung nur jeweils einmalige Datentransfers vorsieht. Zudem ist umstritten, ob der Anspruch auf Datenportabilität sich auch auf beobachtete personenbezogene Daten erstreckt. Auf abgeleitete Daten findet die Norm keine Anwendung. Siehe näher *Janal*, Data portability under the GDPR: A blueprint for access rights?, in: Drexl (Hrsg.), Datenzugang, Verbraucherinteressen und Gemeinwohl, 2021 S. 327 ff.

wenden und personenbezogene Daten Dritter unzulässigerweise zu verarbeiten.[57] Der Betreiber des ADM-Systems kann durchaus ein Interesse daran haben, seine Nutzer:innen vor einer extensiven Datenverarbeitung Dritter zu schützen,[58] zumal einzelne Datensubjekte den Verarbeitungsumfang ihrer Daten und den eventuellen Drittbezug dieser Daten oftmals nicht vollständig erfassen. Selbstverständlich kann ein solches Schutzinteresse auch nur vorgeschoben werden, um eine Analyse des Systems zu unterbinden (was insbesondere dann naheliegt, wenn der Betreiber des ADM-Systems sich weitreichende Einwilligungen zur Datenweitergabe an Dritte einräumen lässt).

Dieses Schutzinteresse des Systembetreibers ist in den AGB mit den Interessen der Nutzer:innen an der Spende oder Monetarisierung ihrer Daten sowie mit den berechtigten Forschungsinteressen in einen angemessenen Ausgleich zu bringen. Dies ließe sich durch Einführung eines Forschungsvorbehalts mit Registrierungspflicht oder einer Klausel mit Erlaubnisvorbehalt erzielen.[59] Ein pauschales Verbot der automatisierten Datenspende stellt meines Erachtens eine unangemessene Benachteiligung i. S. d. § 307 Abs. 1 BGB dar.

4. Bewertung

Das Geschäftsgeheimnisgesetz steht einer unautorisierten Blackbox-Analyse auf Basis von Datenspenden nicht entgegen. Soweit die Nutzung des ADM-Systems den Abschluss eines Nutzungsvertrags voraussetzt, kann die Systemanalyse jedoch gegen die Nutzungsbedingungen verstoßen. Nach hier vertretener Auffassung darf durch Allgemeinen Geschäftsbedingungen zwar das Anlegen fiktiver Nutzerkonten verboten werden, nicht aber die Analyse mittels Datenspenden. Erfolgt die Datenspende in einem automatisierten Prozess, sind Klauseln zulässig, die Registrierpflichten bzw. Erlaubnisvorbehalte enthalten. Zuzugestehen ist allerdings,

---

[57] Der an der Cambridge University beschäftigte Wissenschaftler *Aleksandr Kogan* hatte über die App „thisisyourdigitallife" Persönlichkeitstests angeboten und diese mit dem Unternehmen Cambridge Analytica geteilt. Die App wurde von ca 300.000 Personen genutzt, doch war es Cambridge Analytica aufgrund des von Facebook bereitgestellten API-Zugangs möglich, auch die Daten aller Kontakte auszulesen. Nach Schätzungen wurden die personenbezogenen Daten von über 50 Millionen Facebook-Nutzern an Cambridge Analytica weitergeleitet. Siehe *Segarra*, Mark Zuckerberg Just Revealed 3 Steps Facebook Is Taking to Address the Cambridge Analytica Crisis, 21.3.2018, https://time.com/5209729/mark-zuckerberg-facebook-cambridge-analytica/.

[58] Siehe zu Klagen Facebooks gegen automatisierte Datenextrahierung *Cimpano*, Facebook sues Ukrainian browser extension makers for scraping user data, 10.3.2019, https://www.zdnet.com/article/facebook-sues-ukrainian-browser-extension-makers-for-scraping-user-data/; *Romero*, Combating Scraping by Malicious Browser Extensions, 14.1.2021, https://about.fb.com/news/2021/01/combating-scraping-by-malicious-browser-extensions; *dies.*, Taking Legal Action Against Data Scraping, 1.10.2020, https://about.fb.com/news/2020/10/taking-legal-action-against-data-scraping.

[59] Gegen eine solche Registrierpflicht de lege ferenda *Persily*, Proposal for a Platform Transparency and Accountability Act, sec. 7, https://www.dropbox.com/s/5my9r1t9ifebfz1/Persily%20proposed%20legislation%2010%205%2021.docx?dl=0.

dass insoweit erhebliche Rechtsunsicherheit besteht. Personen, die eine unautorisierte Blackbox-Analyse betreiben, bewegen sich auf ungesichertem Grund und lassen sich gegebenenfalls durch die Androhung gerichtlicher Schritte von Seiten der Entwickler des ADM-Systems abschrecken.[60]

Unabhängig von eventuellen rechtlichen Beschränkungen ist die unautorisierte Blackbox-Analyse aufwändig und stets dem potentiellen Vorwurf ausgesetzt, aufgrund der Auswahl der Datenspender:innen kein zutreffendes Bild des Entscheidungsmechanismus zu zeichnen. Für eine adäquate Kontrolle von ADM-Systemen genügt die Möglichkeit der unautorisierten Blackbox-Analyse nicht. *De lege ferenda* gilt es deshalb, einen Datenzugriff Dritter zu ermöglichen, um eine externe Kontrolle von ADM-Systemen zu ermöglichen.

V. Kontrolloptionen de lege ferenda

1. Behördliche Aufsicht

Als Option der externen Kontrolle ist zunächst die behördliche Aufsicht zu nennen. Gerade mit Blick auf die Vertraulichkeit von Geschäftsgeheimnissen und personenbezogenen Daten weist eine behördliche Kontrolle Vorzüge auf. Geheimhaltungsinteressen lassen sich mittels behördlicher Verschwiegenheitspflichten sowie der Beschränkung von Auskünften nach §§ 5, 6 Informationsfreiheitsgesetz wahren. Entsprechende Regelungen enthalten auch Art. 63 Abs. 4, Abs. 6 des Verordnungsvorschlags für einen Digital Services Act (DSA-E) sowie Art. 70 Abs. 1 lit. a des Verordnungsvorschlags für ein Gesetz über Künstliche Intelligenz (KI-VO-E).

Allerdings verfügen traditionelle Aufsichtsbehörden über Aufgabengebiete, die nicht spezifisch auf die Kontrolle von ADM-Systemen zugeschnitten sind. Ressourcen und Kompetenzen zur Gewährleistung einer effektiven Aufsicht fehlen teilweise. Die Europäische Kommission strebt deshalb die Schaffung neuer Befugnisse und neuer Behörden an. Sowohl Art. 31 Abs. 1 DSA-E als auch Art. 64 KI-VO-E sehen unter bestimmten Voraussetzungen behördliche Aufsichtsbefugnisse zur dynamischen Kontrolle von automatisierten Entscheidungssystemen vor. Damit verbunden ist die Pflicht der betroffenen Unternehmen, den zuständigen Behörden Zugang zu ihren Datensystemen zu gewähren, um eine Überwachung der Vorgaben der jeweiligen Verordnung zu ermöglichen. Art. 64 Abs. 2 KI-VO-E ermöglicht den Behörden erforderlichenfalls auch einen Zugriff auf den Quellcode des ADM-Systems. Im Falle des DSA-E nehmen diese Befugnisse der „Koordinator für digitale Dienste“ sowie die Europäische Kommission wahr,[61] im

---

[60] *Kayser-Bril*, Nach Drohungen von Facebook: AlgorithmWatch sieht sich gezwungen, Instagram-Forschungsprojekt einzustellen, https://algorithmwatch.org/de/instagram-forschung-von-facebook-gestoppt/.

[61] Art. 31 Abs. 1 DSA-E.

KI-VO-E ist die Aufgabe den von den Mitgliedstaaten zu bestimmenden nationalen Marktüberwachungsbehörden zugewiesen.[62] Ein großer Vorteil einer derartigen behördlichen Kontrolle ist die Ergänzung des Datenzugriffs um weitere Überwachungsbefugnisse, wie beispielsweise die Durchsuchung von Räumlichkeiten oder die Aufforderung zur Informationserteilung.[63] Solche Überwachungsbefugnisse werden gegenwärtig in Art. 41, 51 ff. DSA-E anvisiert, während Art. 64 Abs. 3 KI-VO-E die Untersuchungsmöglichkeiten auf die nach dem Verordnungsentwurf zu erstellenden Dokumente beschränkt.

Die in den Verordnungsentwürfen der Kommission vorgesehenen Datenzugriffsmöglichkeiten gelten freilich nur für sehr große Diensteanbieter i. S. d. Art. 25 DSA-E bzw. für Hochrisiko-KI i. S. d. Art. 6 KI-VO-E. In welchem Umfang daneben mitgliedstaatliche Kompetenzen verbleiben, ist derzeit noch Gegenstand der politischen Diskussion im jeweiligen Gesetzgebungsverfahren. Unabhängig von den letztlichen Kompetenzzuweisungen handelt es sich um ein Zukunftsprojekt: Es bedarf zunächst des Aufbaus von Kompetenzen und Ressourcen bei den neu geschaffenen oder neu zuständigen Behörden.

2. Unabhängiges Audit

Als Ergänzung zu einer behördlichen Überwachung kommen Audits unabhängiger Dritter in Betracht. Auch dieses Instrument wird von der Europäischen Kommission in ihren Verordnungsvorschlägen aufgegriffen, allerdings erneut nur für bestimmte Systeme. So sieht Art. 28 DSA-E ein externes Audit für sehr große Online-Plattformen vor, wobei insbesondere auch die für Empfehlungen und Rankings eingesetzten ADM-Systeme erfasst sind.[64] Der Verordnungsvorschlag für einen Digital Markets Act (DMA-E)[65] visiert in Art. 13 ein jährliches Audit der Techniken für Kunden-Profiling an. Die Auditverfahren sind verbunden mit öffentlichen Berichtspflichten, wobei Berichte zum Zwecke des Geheimnisschutzes geschwärzt werden können.[66] Ausgerechnet der Vorschlag für eine KI-VO enthält kein unabhängiges Prüfverfahren, sondern setzt vorwiegend auf eine Eigenbewertung durch die Entwickler:innen.[67]

---

[62] Art. 59 Abs. 1, Abs. 2 Entwurf AI Act.

[63] Näher *BKartA*, Sektoruntersuchung Vergleichsportale, April 2019, https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Sektoruntersuchungen/Sektoruntersuchung_Vergleichsportale_Bericht.pdf?__blob=publicationFile&v=7, S. 139 ff.

[64] Dies folgt aus dem Verweis des Art. 28 Abs. 1 auf die Pflichten sehr großer Online-Plattformen (auch) nach Art. 27 Abs. 1 lit. a DSA-E.

[65] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über bestreitbare und faire Märkte im digitalen Sektor (Gesetz über digitale Märkte), COM/2020/842 final.

[66] Art. 33 Abs. 3 S. 1 DSA-E.

[67] Art. 43 KI-VO-E. Siehe die zutreffende Kritik von *Veale/Zuiderveen Borgesius*, Demystifying the Draft EU Artificial Intelligence Act, CRi 2021, 97, 106.

Externe Prüfungsgesellschaften zur Analyse der Funktionsweise von ADM-Systemen existieren bereits, beispielsweise um ADM-Systeme auf diskriminierende Effekte hin zu kontrollieren.[68] Da das Feld gerade erst im Entstehen begriffen ist, gibt es jedoch gegenwärtig keine einheitlichen Kontrollstandards.[69] Die Nachteile eines Auditverfahrens liegen zudem in den hohen Kosten, welche für die überprüften Marktakteure mit dem Audit verbunden sind sowie in der finanziellen Abhängigkeit der Prüfungsgesellschaften von dem zu kontrollierenden Unternehmen.[70]

3. Forschungszugriff

In Anbetracht der nicht ausreichenden Kontrolle durch Behörden und unabhängige Prüfgesellschaften sowie der gegenwärtig noch unklaren Kontrollmaßstäbe ist es ferner sinnvoll, einen Zugriff zu automatisierten Entscheidungssystemen zugunsten von Wissenschaft und Forschung zu schaffen. Dieser würde es Wissenschaftler:innen ermöglichen, mit unterschiedlichsten Forschungszugängen und aus verschiedenster Perspektive automatisierte Entscheidungssysteme zu analysieren und dabei ihre jeweilige Expertise einzubringen. Der deutsche Gesetzgeber hat dies erkannt und mit § 19 Abs. 3 UrhDaG erstmals einen Anspruch auf Forschungsdatenzugriff gegenüber Betreibern von ADM-Systemen geschaffen. Weniger weit reicht § 5a Abs. 2 Nr. 1 NetzDG, der lediglich einen qualifizierten Auskunftsanspruch enthält.[71] Die Normen haben einen engen Anwendungsbereich: Gegenstand des Auskunftsanspruchs nach § 5a Abs. 2 Nr. 1 NetzDG sind die von sozialen Netzwerken eingesetzten Filter zur Erkennung von Hassrede;

---

[68] Siehe näher *Ng*, Can Auditing Eliminate Bias from Algorithms?, 23.2.2021, https://themarkup.org/ask-the-markup/2021/02/23/can-auditing-eliminate-bias-from-algorithms.

[69] *Schellmann*, Auditors are testing hiring algorithms for bias, but there's no easy fix, 11.2.2021, https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain; *Miller*, Radical Proposal: Third-Party Auditor Access for AI Accountability, 20.10.2021, file:///C:/Users/bt305258/Documents/Aufsätze%20&%20Projekte/Systemische%20Risiken/Radical%20Proposal%20Third-Party%20Auditor%20Access%20for%20AI%20Accountability.htm.

[70] Zu weiteren Risiken siehe *Sloane*, The Algorithmic Auditing Trap, 17.3.2021, https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d. Zu bekannten Problemen aus der Wirtschaftsprüfung *Scholtes*, Neue Regeln für Wirtschaftsprüfer überzeugen nicht, 16.1.2021, https://www.deutschlandfunk.de/nach-dem-wirecard-skandal-neue-regeln-fuer.724.de.html?dram:article_id=490968. Zu daraus resultierenden Gestaltungsempfehlungen für die Kontrolle von Diensteanbietern der Informationsgesellschaft siehe *Schindler*, CEP Policy Brief: Proposed EU Digital Services Act (DSA): Provisions concerning auditing and opportunities to strengthen this mechanism, Juni 2021, https://www.counterextremism.com/sites/default/files/2021-06/CEP_Proposed%20Digital%20Services%20Act_Recommendations%20for%20strengthening%20auditing%20regime_060721.pdf.

[71] Zu den Erwägungen des Gesetzgebers siehe Beschlussempfehlung und Bericht des Ausschusses für Recht und Verbraucherschutz, BT-Drucks. 19/29392; näher *Specht-Riemenschneider*, Studie (Fn. 26), S. 58.

der Forschungsdatenzugang nach § 19 Abs. 3 UrhDaG bezieht sich auf die von Diensteanbietern zum Teilen von Online-Inhalten eingesetzten Filter zur Erkennung urheberrechtswidriger Inhalte. Erfahrungen mit diesen Normen bleiben abzuwarten. Auch der Verordnungsvorschlag für den Digital Services Act enthält in Art. 31 Abs. 2 eine ähnliche Regelung für sehr große Vermittlungsdienste der Informationsgesellschaft.

Je nach Art des ADM-Systems ist der Forschungsdatenzugriff unterschiedlich zu gestalten.[72] So ist zunächst der Kreis der begünstigten Personen zu definieren, wobei zu berücksichtigen ist, dass Forschung nicht allein an Universitäten stattfindet. Gerade bei der Untersuchung von ADM-Systemen haben auch Institutionen der Zivilgesellschaft[73] und gemeinnützige Recherchenetzwerke[74] wertvolle Arbeit geleistet. Gleichzeitig belegt der Cambridge Analytica-Skandal, dass eine universitäre Anbindung nicht vor dem Missbrauch personenbezogener Daten schützt. Eine behördliche Akkreditierung der zugriffsbefugten Forschungsinstitutionen sowie strafbewehrte Vertraulichkeitsgarantien erscheinen deshalb sinnvoll. Zu klären ist auch, auf welche Daten sich der Forschungsdatenzugriff erstrecken soll. Wird der Datenzugriff auf die Daten von innerhalb der EU ansässigen natürliche Personen bzw. Unternehmen beschränkt, hindert dies gegebenenfalls die Erforschung von ausländischen Einflüssen – beispielsweise im Kontext der Desinformation. Außerdem wird auf diesem Wege eine Erforschung von Einflüssen der ADM-Systeme auf Entwicklungsländer eingeschränkt, in denen evtl. Mittel und Expertise zur eigenständigen Analyse solcher Systeme fehlen. Gegen eine Erstreckung des Forschungsdatenzugriffs auf exterritoriale Daten und Ereignisse spricht freilich die Vorbildfunktion eines europäischen Rechtsakts für andere Nationen.[75] Spiegelbildliche Datenzugriffsrechte aus dem Ausland liegen regelmäßig nicht im europäischen Interesse.

Offen ist zudem, für welche Forschungszwecke ein Forschungsdatenzugang gewährt werden sollte. Denn neben der Erforschung von Funktionsweise und Risiken des untersuchten ADM-Systems eröffnen einige dieser Systeme tiefe Einblicke in menschliche Verhaltensweisen bzw. die Mensch-Maschine-Interaktion und schaffen damit günstige Bedingungen für die sozialwissenschaftliche For-

---

[72] Eingehend *Specht-Riemenschneider*, Studie (Fn. 26).

[73] Z. B. die Untersuchungen von Algorithmwatch zum Schufa-Score und Instagram-Algorithmus, näher https://algorithmwatch.org.

[74] Die bekannte COMPAS-Untersuchung geht auf die gemeinnützige Nachrichtenredaktion ProPublica zurück, *Angwin/Larson/Mattu/Kirchner*, Machine Bias – There's software used across the country to predict future criminals. And it's biased against blacks, 23.5.2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[75] Zur internationalen Vorbildfunktion des NetzDG siehe *Mchangama/Alkiviadou*, The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype For Global Online Censorship – Act Two, September 2020, https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf.

schung. Schließlich gilt es zu klären, in welchem Umfang die Ergebnisse dieser Forschung publiziert werden dürfen und auf welche Weise personenbezogene Daten und vertrauliche Unternehmensdaten im Publikationsprozess zu schützen sind.

VI. Fazit

Von Privaten betriebene automatisierte Entscheidungssysteme werden zunehmend eingesetzt, um Entscheidungen zu treffen oder Empfehlungen auszusprechen, die nicht nur die von dieser Entscheidung betroffene Person tangieren. In der Summe haben diese Entscheidungen bzw. Empfehlungen auch erhebliche gesellschaftliche Implikationen. Regelmäßig sind diese ADM-Systeme nicht von externen Stellen validiert. Die Daten, Modellierungsentscheidungen und Software, die den Entscheidungen zugrunde liegt, werden unter Berufung auf den Daten- und Geschäftsgeheimnisschutz oftmals nicht offengelegt. Gegenwärtig existierende Informations- und Offenlegungspflichten sind sektorspezifisch und oberflächlich. Sie sind weder geeignet noch dazu konzipiert, eine externe Kontrolle der Entscheidungsmechanismen zu ermöglichen. Versuche, die als „Blackbox" empfundenen ADM-Systeme von außen mittels Datenspenden und fiktiven Nutzerkonten zu analysieren, verstoßen nach hier vertretener Auffassung zwar nicht gegen das Geschäftsgeheimnisgesetz und können durch AGB nur begrenzt unterbunden werden. Doch ist die Rechtslage unsicher, was einen abschreckenden Effekt auf die Analyst:innen haben kann. Eine Klarstellung durch den Gesetzgeber wäre wünschenswert.

Ergänzend bedarf es neuer Instrumente, um die Entscheidungsqualität, Sicherheit und gesellschaftlichen Folgen automatisierter Entscheidungssysteme einer externen Kontrolle zu unterziehen. Neben zusätzlichen behördlichen Befugnissen und unabhängigen Auditverfahren kommen hier namentlich Forschungsdatenzugänge in Betracht.[76] Der deutsche Gesetzgeber hat mit § 19 Abs. 3 UrhDaG und § 5a NetzDG einen ersten wichtigen Aufschlag gemacht. Auch dem Unionsgesetzgeber ist die Problematik ausweislich der Art. 31 Abs. 2 DSA-E, Art. 13 DMA bewusst. Ausgerechnet der Entwurf für eine KI-VO bleibt in dieser Hinsicht zu blass. Letztlich wird eine Kombination der genannten Instrumente erforderlich sein, um Licht in das Dunkel der Funktionsweise automatisierter Entscheidungssysteme zu bringen, die von privater Hand betrieben werden.

[76] Jeweils sind die existierenden Erlaubnistatbestände für die hiermit verbundene Verarbeitung personenbezogener Daten zu prüfen und ggf. zu ergänzen, hierzu *Specht-Riemenschneider*, Studie (Fn. 26), S. 34.

Arbeitsmarktchancen per Algorithmus?

Zur Allokation von Maßnahmen der aktiven Arbeitsmarktpolitik durch algorithmische Entscheidungssysteme

Rüdiger Krause[1]

I. Einleitung

Im Zuge der fortschreitenden Digitalisierung gewinnen algorithmische Systeme bei Entscheidungsvorgängen in den verschiedensten Lebensbereichen immer mehr an Bedeutung.[2] So wird eine wachsende Anzahl von Entscheidungsprozessen in der Weise strukturiert, dass ein informationstechnisches System maschinell trainierter Algorithmen (Modell)[3] einen personenbezogenen Datensatz (Input) auswertet,[4] um hieraus neue Daten (Output) zu gewinnen, die anschließend als Grundlage von für die betroffenen Personen nicht trivialen Entscheidungen dienen.[5] Die Brisanz dieser neuartigen Entscheidungsarchitektur ergibt sich somit nicht allein aus der Komplexität der durchgeführten Rechenoperationen[6] oder aus der schieren Menge der verarbeiteten Daten (Big Data),[7] sondern vor allem

---

[1] Der Verfasser dankt Herrn wissenschaftlichen Mitarbeiter *Jonas Walter Kühn* für die wertvolle Unterstützung. Soweit im nachfolgenden Text für Personen das generische Maskulinum verwendet wird, sind damit alle geschlechtlichen Identitäten und Zuordnungen gemeint.

[2] *Hoffmann-Riem*, AöR 142 (2017), 1 (3 ff.); *Martini*, JZ 2017, 1017 (1017); *Wischmeyer*, AöR 143 (2018), 1 (2); siehe auch Wissenschaftliche Dienste des Deutschen Bundestages, Einsatz und Einfluss von Algorithmen auf das digitale Leben, Nr. 26/17 (2017); aus internationaler Perspektive prägnant *Danaher*, Philosophy & Technology 29 (2016), 245 (245): „age of algorithmic decision-making"; ferner *Olhede/Wolfe*, Philosophical Transactions, Royal Society A 376 (2018), 1 (1 f.); *Zarsky*, Science, Technology & Human Values 41 (2016), 118 (119).

[3] Zu Algorithmen allgemein *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz (2019), S. 17 ff.; aus der Sicht der Informatik umfassend *Sedgewick/Wayne*, Algorithmen: Algorithmen und Datenstrukturen, 4. Aufl. (2014); analytisch ferner *Dourish*, Big Data & Society 3 (2) (2016), 1 (3 ff.).

[4] Regelmäßig durch Priorisieren, Klassifizieren, Sortieren und Filtern, vgl. etwa *Kolleck/Orwat*, Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick, TAB-Hintergrundpapier Nr. 24 (2020), S. 20; siehe auch Datenethikkommission, Gutachten (2019), S. 62, 160; Konrad Adenauer Stiftung (Hrsg.), Algorithmische Entscheidungen: Transparenz und Kontrolle, Analysen & Argumente Nr. 338 (2019), S. 2 ff.

[5] Zu den Begrifflichkeiten siehe auch *Wischmeyer*, AöR 143 (2018), 1 (4 Fn. 9).

[6] Kritisch zur Fokussierung auf Algorithmen deshalb *Wischmeyer*, AöR 143 (2018), 1 (4 Fn. 9).

[7] Zum Big Data-Diskurs statt vieler *Knorre/Müller-Peters/Wagner*, Die Big-Data-Debatte (2020).

daraus, dass die zur Anwendung kommenden Algorithmen von privaten oder öffentlichen Akteuren zur Verfolgung bestimmter Ziele und Interessen in sozial relevante Entscheidungssituationen eingebettet werden.[8] Hierbei sind die Auswirkungen der algorithmischen Systeme innerhalb des jeweiligen Handlungsfeldes[9] umso größer, je geringer der menschliche Einfluss auf die letztlich getroffenen Entscheidungen ist.[10] Von besonderer Relevanz sind in diesem Zusammenhang solche Systeme, die nicht ausschließlich aus regelbasierten Algorithmen bestehen, sondern in denen auch nicht regelbasierte („lernende") Algorithmen zum Einsatz kommen, die sich mit den zugeführten Daten gleichsam selbst programmieren, um durch kontinuierliche Feedback-Schleifen zu immer präziseren Ergebnissen häufig mit dem Ziel von Mustererkennungen zu gelangen (Maschinelles Lernen).[11] Dabei mag es hier dahinstehen, ab welchem Reifegrad der technischen Entwicklung man bei algorithmischen Systemen von „Künstlicher Intelligenz"[12] sprechen kann,[13] indem ein Stadium erreicht ist, bei dem sich das System als „Black Box" darstellt,[14] weil der konkrete Weg zum jeweiligen Output selbst von Experten nicht mehr umfassend nachvollzogen werden kann[15].

Zu denjenigen Lebensbereichen, in die algorithmische Entscheidungssysteme immer weiter vordringen, gehört auch die Arbeitswelt. Hierbei geht es zum einen – neben betrieblichen Systemen zur Steuerung und Kontrolle des unmittelbaren Arbeitsprozesses (Algorithmic Management) – zunehmend um eine datengestützte

---

[8] *Herzog*, DZPhil 69 (2021), 197 (201); *Mittelstadt* et al., Big Data & Society 3 (2) (2016), 1 (2 f.); *Wischmeyer*, AöR 143 (2018), 1 (4 Fn. 9). Siehe auch *Kitchin*, Information, Communication & Society 20 (2017), 14 (25 f.). Die Rolle von Menschen bei der Frage, *welches* Wissen durch Daten produziert werden soll, betont zu Recht *Lopez*, Merkur 863 (2021), 42 (45). Zum menschlichen Faktor bei der Entstehung algorithmischer Systeme ferner *Hoffmann-Riem*, AöR 145 (2020), 1 (12).

[9] Dazu generell *Ananny*, Science, Technology & Human Values 41 (2016), 93 (97 f.); *Beer*, Information, Communication & Society 20 (2017), 1 (4 ff.).

[10] Insoweit kann mit der Datenethikkommission klassifikatorisch zwischen algorithmenbasierten, algorithmengetriebenen und algorithmendeterminierten Entscheidungen differenziert werden, vgl. Datenethikkommission, Gutachten (2019), S. 24, 161.

[11] Zu den unterschiedlichen Ansätzen und Methoden anschaulich Fraunhofer-Gesellschaft, Maschinelles Lernen (2018), S. 9 ff.; ferner *Bilski/Schmid*, NJOZ 2019, 657 (657 ff.); *Linardatos*, Autonome und vernetzte Aktanten im Zivilrecht (2021), S. 52 ff.

[12] Hierzu *Ertel*, Grundkurs Künstliche Intelligenz, 5. Aufl. (2021), S. 1 ff.; *Niederrée/Nejdl*, in: Ebers et al. (Hrsg.), Künstliche Intelligenz und Robotik (2020), § 2 Rn. 1 ff.; *Stiemerling*, in: Kaulartz/Braegelmann (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning (2020), Kapitel 2.1 Rn. 6 ff.

[13] Eine Definition von „Künstlicher Intelligenz" findet sich nunmehr in Art. 3 Nr. 1 i. V. m. Anhang I des Vorschlags der Europäischen Kommission für eine KI-Verordnung, COM(2021) 206 final, S. 46. Dazu *Spindler*, CR 2021, 361 (362 f.). Zur Begrifflichkeit erhellend auch *Herberger*, NJW 2018, 2825 (2825 ff.).

[14] Grdl. *Pasquale*, The Black Box Society (2015).

[15] Insoweit wird häufig auf den Einsatz sog. neuronaler Netze abgestellt; so etwa *Bilski/Schmid*, NJOZ 2019, 657 (659); *Höpfner/Daum*, ZFA 2021, 467 (471 f.); *Huff/Götz*, NZA Beilage 2/2019, 73 (76); siehe auch *Stalder*, Kultur der Digitalität, 5. Aufl. (2021), S. 177 ff.; skeptisch gegenüber dieser Einschätzung unter Verweis auf ältere Aussagen über den Black Box-Charakter von komplexer Software *Passig*, Merkur 823 (2017), 16 (18 ff.).

Personalverwaltung (Human Resource Management)[16] entlang des gesamten Mitarbeiterpfades von der Rekrutierung bis zum Ausscheiden aus dem Unternehmen.[17] Zum anderen sind Systeme zu nennen, mit denen die öffentlichen Arbeitsverwaltungen ihre Tätigkeit effektiver gestalten und insbesondere dem Phänomen von Langzeitarbeitslosigkeit entgegenwirken wollen.[18] Dazu zählt das im Auftrag des österreichischen Arbeitsmarktservice (AMS) entwickelte Arbeitsmarktchancen-Assistenz-System (AMAS), das unter dem Schlagwort „AMS-Algorithmus" national[19] wie international[20] für erhebliches Aufsehen gesorgt und in Österreich sogar eine regelrechte Kampagne mit dem Ziel ausgelöst hat, den Einsatz dieses Systems aufzuhalten („Stoppt den AMS-Algorithmus").[21] Im Kern läuft dieses algorithmische System, das als ein besonders prominenter Fall im Zentrum der folgenden Ausführungen stehen soll, darauf hinaus, die für den einzelnen Arbeitsuchenden bedeutsame Förderpraxis des AMS zumindest bis zu einem gewissen Grade an der mithilfe von Algorithmen errechneten Zuordnung zu bestimmten Gruppen von Arbeitsuchenden auszurichten.

Das aufgeworfene Thema wird nachfolgend in vier Schritten entfaltet: Zum besseren Verständnis sollen zunächst die Funktionsweise und der Kontext des „AMS-Algorithmus" näher vorgestellt werden (II.). Sodann geht es vor dem Hintergrund der bereits schwebenden juristischen Auseinandersetzungen um die spezifisch datenschutzrechtlichen Gesichtspunkte beim Einsatz dieses algorithmischen Systems (III). Weitere Überlegungen gelten den sonstigen Risiken, namentlich der Gefahr von Diskriminierungen, die mit der algorithmenbasierten Berechnung von Arbeitsmarktchancen und einer darauf fußenden Auswahl von Förder- und Betreu-

---

[16] Zur Entwicklung dieses HR-Zweiges instruktiv *Marler/Boudreau*, International Journal of Human Resource Management 28 (2017), 3 ff.

[17] Überblick bei *Gießler*, Was ist automatisiertes Personalmanagement? (2021); siehe auch *Dahm/Dregger*, in: Hermeier/Heupel/Fichtner-Rosada (Hrsg.), Arbeitswelten der Zukunft (2019), S. 249 ff.; *Freyler*, NZA 2020, 284 ff.; *Greif/Kullmann*, ZAS 2021, 61 ff.; *Höpfner/Daum*, ZFA 2021, 467 ff.; *Huff/Götz*, NZA Beilage 2/2019, 73 ff.; *Imping*, DB 2021, 1808 ff.; *Jaspers/Jacquemain*, RDV 2019, 232 ff.; *Joos*, NZA 2020, 1216 ff.; *Malorny*, RdA 2022, 170 ff.; *Prassl*, DRdA 2022, 195 ff.; umfassende Darstellung bei *Petry/Jäger* (Hrsg.), Digital HR (2018); monographisch *Blum*, People Analytics (2021); *Götz*, Big Data im Personalmanagement (2020). Aus internationaler Perspektive *Bernhardt/Kresge/Suleiman*, Data and Algorithms at work (2021); ferner *Kellogg/Valentine/Christin*, Academy of Management Annals 14 (2020), 366 ff., sowie die zahlreichen unter dem Titel „Automation, Artificial Intelligence, and Labor Law" in Comparative Labor Law & Policy Journal 41 (2019) versammelten Beiträge. Anschauliche Systematisierung datenverarbeitender Systeme im Beschäftigungskontext bei *Christl*, Digitale Überwachung und Kontrolle am Arbeitsplatz (2021), S. 27 f.

[18] Vgl. *Desiere/Langenbucher/Struyven*, Statistical profiling in public employment services: An international comparison, OECD Social, Employment and Migration Working Papers No. 224 (2019); *Scoppetta/Buckenleib*, Tackling Long-Term Unemployment through Risk Profiling and Outreach, ESF Transnational Platform, Technical Dossier no. 6 (2018).

[19] Vgl. *Allhutter*, WISO 1/2021, 81 ff.; *Allhutter* et al., Der AMS-Algorithmus, Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften (2020); *Gerhartl*, ZIIR 2021, 24 ff.; *Thiele*, ZIIR 2020, 410 ff.; *Wagner* et al., juridikum 2020, 191 ff.

[20] Vgl. *Büchner/Dosdall*, Köln Z Soziol 73 (2021), Suppl 1, 333 ff.

[21] Siehe https://amsalgorithmus.at.

ungsangeboten einhergehen (IV.). Den Abschluss bilden Überlegungen zu Lösungsansätzen, mit denen die Vorzüge eines solchen algorithmischen Systems genutzt werden können, die damit verbundenen Risiken für die unmittelbar Betroffenen sowie für die Gesellschaft insgesamt aber möglichst gering gehalten werden (V.).

II. Funktionsweise und Kontext des „AMS-Algorithmus"

Der österreichische Arbeitsmarktservice ist ein Unternehmen des öffentlichen Rechts, das mit der Durchführung der österreichischen Arbeitsmarktpolitik nach Maßgabe des Arbeitsmarktservicegesetzes (AMSG)[22] betraut ist und die arbeitsmarktpolitischen Ziele weitgehend unabhängig von ministeriellen Vorgaben des zuständigen Bundesministers für Arbeit und Soziales umsetzt.[23] Nach bis in das Jahr 2008 zurückreichenden Vorläufern beauftragte der AMS im Jahr 2015 ein privates Unternehmen mit der Entwicklung eines „Arbeitsmarktchancen-Assistenz-Systems", mit dem die Chancen der Integration der einzelnen Arbeitsuchenden in den österreichischen Arbeitsmarkt anhand einer Vielzahl von statistischen Daten errechnet werden sollen, um die Förder- und Betreuungsformate des AMS an den auf diese Weise ermittelten Integrationschancen zu orientieren. Im Herbst 2018 ging das AMAS in den Testbetrieb. Die landesweite Implementierung war – nach mehrmaligen Verschiebungen – für Januar 2021 vorgesehen. Eine vorherige Intervention der österreichischen Datenschutzbehörde sowie ein sich daran anschließender und bislang noch nicht beendeter Rechtsstreit[24] haben allerdings dazu geführt, dass der Einsatz des AMAS aktuell ausgesetzt ist.

Mit dem AMAS werden drei übergreifende Ziele verfolgt: Erstens soll die Effizienz und Effektivität der Beratung seitens des AMS erhöht werden. Zweitens sollen die arbeitsmarktpolitischen Fördermaßnahmen gezielter eingesetzt und auf diejenigen Arbeitsuchenden konzentriert werden, bei denen die größte Effektivität der Maßnahmen zu erwarten ist. Drittens schließlich soll die Vergabe von Fördermitteln standardisiert und einer Willkür bei Entscheidungen seitens der Sachbearbeiter beim AMS entgegengewirkt werden.

Zu diesem Zweck wurden in einem ersten Schritt drei Kategorien von Arbeitsuchenden gebildet, denen das System jeweils unterschiedliche Chancen der Integration in den Arbeitsmarkt zuordnet (hohe, mittlere und niedrige Chancen). Von einer hohen Integrationschance geht das System dann aus, wenn eine

---

[22] Bundesgesetz über das Arbeitsmarktservice (Arbeitsmarktservicegesetz – AMSG), BGBl. Nr. 313/1994 i. d. F. BGBl. I Nr. 86/2022 (Österreich); aktuelle Fassung des AMSG abrufbar unter https://www.ris.bka.gv.at.

[23] Zum Folgenden ausführlich *Allhutter*, WISO 1/2021, 81 ff.; *Allhutter* et al. (Fn. 19); *Büchner/Dosdall*, Köln Z Soziol 73 (2021), Suppl 1, 333 (337 ff.); *Lopez*, in: Conference proceedings of the 18th STS Conference Graz 2019 (2019), S. 289 ff.; zum technischen Hintergrund siehe auch *Holl/Kernbeiß/Wagner-Pinter*, Das AMS-Arbeitsmarktchancen-Modell (2018).

[24] Dazu sogleich unter sub III 1.

kurzfristige Eingliederung in den Arbeitsmarkt mit einer Wahrscheinlichkeit von mehr als 66 % zu erwarten ist. Eine niedrige Integrationschance liegt dagegen dann vor, wenn die Wahrscheinlichkeit selbst einer langfristigen Eingliederung in den Arbeitsmarkt bei weniger als 25 % liegt. Alle anderen Wahrscheinlichkeitsgrade werden dem mittleren Segment zugewiesen. Die abstrakte Bildung dieser drei Gruppen erfolgt anhand einer Vielzahl von Parametern wie insbesondere Geschlecht, Altersgruppe, Herkunft, Ausbildung, gesundheitlichen Beeinträchtigungen, Betreuungspflichten und dem regionalen Arbeitsmarktgeschehen. Anhand dieser Daten werden für verschiedene Konstellationen von Merkmalen auf der Basis der dem AMS vorliegenden Geschäftsfälle die Integrationserfolge in den Arbeitsmarkt nach vorausgegangener Arbeitslosigkeit für die vergangenen vier Jahre analysiert, um hieraus entsprechende Wahrscheinlichkeitswerte zu errechnen. Der AMS-Algorithmus zielt somit darauf ab, die Frage zu beantworten, mit welcher statistischen Wahrscheinlichkeit eine Person, auf die bestimmte Merkmale zutreffen, in der Vergangenheit nach vorhergehender Arbeitslosigkeit wieder in den Arbeitsmarkt integriert wurde. In einem zweiten Schritt wird der konkrete Arbeitsuchende auf der Grundlage seiner spezifischen Merkmale durch das System in eine der drei Gruppen einsortiert, wodurch ihm ein bestimmter Grad der Wahrscheinlichkeit einer zukünftigen erfolgreichen Arbeitsmarktintegration attestiert wird. Dabei zeigen die bisherigen Erfahrungen mit dem AMS-Algorithmus, dass etwa 4 % der Arbeitsuchenden dem höchsten, 65 % dem mittleren und 31 % dem niedrigsten Segment zugeordnet werden.

Wie eingangs bereits hervorgehoben, darf sich eine Würdigung nicht auf die technischen Funktionalitäten der angewendeten Algorithmen beschränken. Vielmehr bedarf es einer Betrachtung des gesamten soziotechnischen Systems und damit insbesondere der Einbettung der Algorithmen in die Entscheidungsstrukturen des AMS. Insoweit ist zunächst festzuhalten, dass die Vermittlungsaktivitäten des AMS weiterhin allen Arbeitsuchenden gleichermaßen zugutekommen. Dagegen sollen die Förderleistungen für die berufliche Weiterbildung auf das mittlere Segment konzentriert werden, weil sich der AMS hiervon – gemessen an den Integrationserfolgen – die höchste Effektivität der eingesetzten Mittel verspricht. Bei Arbeitsuchenden mit hohen Integrationschancen soll der Schwerpunkt stattdessen auf die Vermittlung gelegt werden, da eine Eingliederung in den Arbeitsmarkt auch ohne eine weitere Förderung zu erwarten sei. Arbeitsuchende mit niedrigen Integrationschancen wiederum sollen dem neu entwickelten externen Format „Beratungs- und Betreuungseinrichtung Neu (BBEN)" zugewiesen werden, um den bei ihnen bestehenden besonderen Vermittlungshindernissen besser Rechnung tragen zu können und eine Fokussierung der Beratungstätigkeit der AMS-Mitarbeiter auf das mittlere Segment zu ermöglichen.[25]

[25] Dazu AMS/prospect (Hrsg.), Evaluierung des Betreuungsformates für Personen mit multiplen Vermittlungshindernissen (BBEN) (2019); siehe auch AMS/WIFO (Hrsg.), Kosten-Ertrags-

Soweit es nun um die konkrete Einbeziehung des AMS-Algorithmus auf der Ebene der Individualbetreuung von Arbeitsuchenden geht, wird der Output, d. h. die errechnete Zuordnung des Betroffenen zu einer der drei Gruppen, dem AMS-Mitarbeiter vom System zu Beginn der jeweiligen Sachbearbeitung angezeigt.[26] Zwar trifft die letztgültige Entscheidung über den Zuschnitt einer etwaigen Förderung der jeweilige AMS-Sachbearbeiter, wobei die internen Richtlinien des AMS die Mitarbeiter durchaus dazu anhalten, je nach der konkreten Sachlage abweichend vom Vorschlag des algorithmischen Systems zu entscheiden. Allerdings ist die Eingabemaske so konfiguriert, dass eine abweichende Beurteilung einen zusätzlichen Begründungsaufwand auslöst, während die Übernahme der Systemvorgabe keine weiteren Maßnahmen erfordert.[27]

III. Datenschutzrechtliche Aspekte

Der AMS-Algorithmus kann aus unterschiedlichen Perspektiven betrachtet werden, wobei im Folgenden vor dem Hintergrund der juristischen Auseinandersetzungen um das AMAS zunächst ein spezifisch datenschutzrechtlicher Blickwinkel gewählt werden soll.

1. Der Rechtsstreit um den AMS-Algorithmus

Wie bereits angedeutet, setzte die österreichische Datenschutzbehörde dem Einsatz des AMAS zumindest ein vorläufiges Ende, indem sie dem AMS mit Bescheid vom 16. August 2020 die weitere Nutzung dieses Systems ab dem 1. Januar 2021 grundsätzlich untersagte.[28] Der gegen diesen Bescheid vom AMS eingelegten Beschwerde gab das österreichische Bundesverwaltungsgericht (BVwG) mit Entscheidung vom 12. Dezember 2020 statt und hob den Bescheid auf.[29] Über die hiergegen von der Datenschutzbehörde an den österreichischen Verwaltungsgerichtshof (VwGH) eingelegte Revision ist bislang nicht entschieden. Allerdings hat der AMS erklärt, bis zum rechtskräftigen Abschluss des Verfahrens von einem weiteren Einsatz des AMAS abzusehen.[30]

Analyse der „Beratungs- und Betreuungsleistungen für Personen mit multiplen Vermittlungshindernissen" (BBEN) (2020).

[26] Vgl. *Allhutter* et al. (Fn. 19), S. 67 f.

[27] Vgl. *Allhutter* et al. (Fn. 19), S. 68 f.

[28] Datenschutzbehörde v. 16.8.2020 – DSB-D213.1020, 2020-0.513.605; vgl. DSB Newsletter 4/2020, 3 f.

[29] BVwG v. 18.12.2020 – W256 2235360 – 1/5 E – ECLI:AT:BVWG:2020:W256.2235360.1.00.

[30] Zum Verfahrensstand siehe die Antwort des österreichischen Bundesministeriums für Arbeit v. 24.8.2021, 7065/AB auf die parlamentarische Anfrage unter der Nr. 7142/J (XXVII. GP).

Im Zentrum des konkreten Rechtsstreits steht die Frage nach dem Vorhandensein einer hinreichenden Rechtsgrundlage für die mit der Verwendung des AMAS verbundene Verarbeitung personenbezogener Daten von Arbeitsuchenden. Aus der Sicht der Datenschutzbehörde stellt der die Datenverarbeitung durch den AMS allgemein regelnde § 25 Abs. 1 AMSG in Verbindung mit den Vorschriften über die Ziele und Aufgaben (§ 29 AMSG) sowie die Grundsätze bei der Aufgabenerfüllung (§ 31 Abs. 5 AMSG) keine ausreichende rechtliche Grundlage dar. Für diese Auffassung fährt die Datenschutzbehörde zwei Argumentationslinien auf: So gehe es beim AMS-Algorithmus der Sache nach um ein Profiling der Arbeitsuchenden, bei dem angesichts des hierdurch gewonnenen „informationellen Mehrwerts" nicht zuletzt vor dem Hintergrund von Art. 4 Nr. 4, Art. 6 Abs. 1 UAbs. 1 Buchst. c und e sowie Art. 9 Abs. 2 Buchst. h DSGVO eine hinreichend präzise Grundlage im mitgliedstaatlichen Recht erforderlich sei, der eine schlichte Auflistung der Daten in § 25 Abs. 1 AMSG, die verarbeitet werden dürfen, nicht genüge. Darüber hinaus sei nicht auszuschließen, dass die Ergebnisse des AMAS von den AMS-Mitarbeitern routinemäßig übernommen würden. Daher lägen insoweit automatisierte Entscheidungen im Sinne von Art. 22 Abs. 1 DSGVO vor, sodass die zusätzlichen Vorgaben dieser Regelung gelten würden, wobei insbesondere die Anforderungen des Art. 22 Abs. 4 DSGVO durch die aktuell vorhandenen österreichischen Vorschriften erst recht nicht erfüllt seien.

Das BVwG ist dieser Argumentation im Wesentlichen mit drei Überlegungen entgegengetreten: Erstens stelle § 25 AMSG im Hinblick auf die Verarbeitung personenbezogener Daten als solche eine hinreichend bestimmte mitgliedstaatliche Vorschrift im Sinne von Art. 6 Abs. 1 UAbs. 1 Buchst. e sowie Art. 9 Abs. 2 Buchst. g DSGVO dar, wobei die Gewährleistung eines möglichst geordneten und gut funktionierenden Arbeitsmarktes im erheblichen öffentlichen Interesse liege. Dementsprechend bestünden „überhaupt keine Bedenken", dass die Daten für die Bewertung der Arbeitsmarktchancen von Arbeitsuchenden verwendet werden dürften. Zweitens würde sich aus der DSGVO nicht ergeben, dass das Profiling für sich genommen einen „informationellen Mehrwert" erzeuge und deshalb höhere Anforderungen an die Konkretheit der mitgliedstaatlichen Rechtsgrundlage zu stellen seien. Drittens schließlich bestünden infolge des vorgesehenen Einsatzes des AMAS als bloßes Hilfsmittel für die AMS-Berater „überhaupt keine Gründe", von einer gänzlich automatisierten Entscheidung im Sinne von Art. 22 Abs. 1 DSGVO auszugehen, weil es insoweit nur auf den Verarbeitungsvorgang als solchen, nicht aber auf etwaige Verstöße gegen organisationsinterne Vorgaben ankomme.

2. Notwendigkeit einer spezifizierten Rechtsgrundlage?

Im Ausgangspunkt steht außer Zweifel, dass im Rahmen des AMAS personenbezogene Daten verarbeitet werden und es sich hierbei um ein Profiling im Sinne

des Art. 4 Nr. 4 DSGVO (in Gestalt eines Dreiphasensystems)[31] handelt. Genauer gesagt erfolgt die Verarbeitung personenbezogener Daten jedenfalls in der Inferenzphase (Anwendungsphase), in welcher der AMS-Algorithmus (d. h. das zuvor abstrakt entwickelte Modell der Arbeitsmarktchancenberechnung) auf die konkreten Merkmale des Arbeitsuchenden angewendet wird, um auf diese Weise Erkenntnisse über die Arbeitsmarktchancen der einzelnen arbeitsuchenden Person zu gewinnen.[32] Bei der Wahrscheinlichkeit einer erfolgreichen Integration in den Arbeitsmarkt geht es zwar um ein Phänomen, das nicht ausschließlich vom Verhalten des Betroffenen abhängt, sondern bei dem auch die antizipierten Reaktionen Dritter einbezogen werden. Wenn die DSGVO die automatisierte Verarbeitung personenbezogener Daten zur Vorhersage der Arbeitsleistung als Profiling bezeichnet, ist ein entsprechendes Verfahren zur Prognose von Arbeitsmarktchancen aber ebenso hierunter zu fassen, erfolgt doch auch in diesem Fall eine Vorhersage bestimmter persönlicher Aspekte.

Allein aus der Einordnung des AMAS als Profiling lassen sich freilich noch nicht die von der Datenschutzbehörde gezogenen Schlussfolgerungen ableiten. So ist zunächst festzuhalten, dass die DSGVO ein Profiling keineswegs untersagt, sondern nur (klarstellend) in den Anwendungsbereich der DSGVO zieht und im Übrigen spezifische Rechtmäßigkeitsanforderungen für den Fall aufstellt, dass ein automatisiertes Profiling ohne weiteren Zwischenschritt in eine automatisierte Entscheidung im Sinne von Art. 22 Abs. 1 DSGVO mündet.[33] Weiter gibt die DSGVO nicht zu erkennen, dass die allgemeinen Rechtfertigungstatbestände gemäß Art. 6 Abs. 1 DSGVO und damit auch Art. 6 Abs. 1 UAbs. 1 Buchst. e DSGVO sowie – bei einer Verarbeitung von gesundheitsbezogenen Daten – Art. 9 Abs. 1 Buchst. g DSGVO als Ausgangspunkt für eine rechtmäßige Verarbeitung personenbezogener Daten bei einem Profiling nicht genügen sollen. Vielmehr spricht Art. 21 Abs. 1 S. 1 Halbs. 2 DSGVO sogar dafür, dass ein Profiling auf Art. 6 Abs. 1 UAbs. 1 Buchst. e DSGVO gestützt werden kann. In ErwGr 72 S. 1 DSGVO heißt es dementsprechend ausdrücklich, dass das Profiling den Vorschriften der Verordnung unterliegt. Richtig ist allerdings, dass Art. 6 Abs. 1 UAbs. 1 Buchst. e und Art. 9 Abs. 2 Buchst. g DSGVO für sich genommen noch keine Legitimationsgrundlage für eine Datenverarbeitung liefern, sondern dass es gemäß Art. 6 Abs. 3 DSGVO einer „Scharniernorm“[34]

---

[31] Zum Dreiphasensystem – bestehend aus Datenaggregation (1. Phase), Modellierung (2. Phase) und Inferenz (3. Phase) – anschaulich *Lorentz*, Profiling – Persönlichkeitsschutz durch Datenschutz? (2020), S. 35 ff., 41 ff.

[32] Ob es zu einer Verarbeitung personenbezogener Daten auch in der Modellierungsphase gekommen ist, hängt davon ab, ob insoweit nicht nur anonymisierte Daten verwendet worden sind, vgl. *Lorentz* (Fn. 31), S. 97 f.

[33] Simitis/Hornung/Spiecker gen. Döhmann/*Scholz*, Datenschutzrecht (2019), Art. 4 Nr. 4 DSGVO Rn. 10. Zur Unterscheidung zwischen Profiling und automatisierter Entscheidung deutlich Kühling/Buchner/*Buchner*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 4 Nr. 4 Rn. 1.

[34] Vgl. Simitis/Hornung/Spiecker gen. Döhmann/*Roßnagel*, Datenschutzrecht (2019), Art. 6 Abs. 1 DSGVO Rn. 71.

bedarf, die in der Tat hinreichend bestimmt sein muss[35]. Sofern für die Verarbeitung personenbezogener Daten als solche eine derartige mitgliedstaatliche Rechtsvorschrift vorhanden ist, lässt sich der DSGVO indes nicht entnehmen, dass ein Profiling in der fraglichen Norm explizit genannt werden muss, um den unionsrechtlichen Vorgaben zu genügen. Im Gegenteil: Wenn ErwGr 71 S. 3 DSGVO eine ausdrückliche rechtliche Grundlage nur für den Fall verlangt, dass ein Profiling die Basis einer automatisierten Entscheidungsfindung ist, spricht dies im Umkehrschluss dafür, dass bei einem „einfachen" Profiling ohne eine anschließende automatisierte Entscheidung gerade keine spezifizierte Rechtsgrundlage erforderlich ist.

3. Vorliegen einer automatisierten Entscheidung?

Von der soeben angesprochenen Frage besonderer Anforderungen an die Bestimmtheit der Rechtsgrundlage bei einem „einfachen" Profiling strikt zu unterscheiden ist das weitere Problem, ob das AMAS angesichts seiner konkreten Ausgestaltung unter die für automatisierte Entscheidungen geltende Regelung des Art. 22 Abs. 1 DSGVO fällt,[36] sodass strengere Rechtmäßigkeitsanforderungen gelten. In diesem Zusammenhang kann vorab festgehalten werden, dass Art. 22 Abs. 1 DSGVO trotz seines Wortlauts, der auf einen Unterlassungsanspruch hindeutet, der vom Betroffenen individuell geltend gemacht werden muss, ein grundsätzliches Verbot von Entscheidungen enthält, die ausschließlich auf einer automatisierten Verarbeitung personenbezogener Daten beruhen und gegenüber dem Betroffenen rechtliche Wirkung entfalten oder in ähnlicher Weise erheblich beeinträchtigen.[37] Dieses schon in Art. 15 Abs. 1 RL 95/46/EG enthaltene prinzipielle Verbot, das vor dem Hintergrund einer entsprechenden Regelung im französischen Recht seinerzeit auf Betreiben Frankreichs in die Richtlinie aufgenommen worden war,[38] soll verhindern, dass Menschen zu bloßen Objekten automatisierter Entscheidungsvorgänge gemacht werden.[39] Anstelle einer Unterordnung von individuellen Menschen unter rein maschinelle

---

[35] Kühling/Buchner/*Buchner*/*Petri*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 6 Rn. 114.

[36] Zur Unterscheidung zwischen der Rechtmäßigkeit eines Profiling und der von Art. 22 DSGVO allein adressierten Form der Nutzung der Ergebnisse der Datenverarbeitung Kühling/Buchner/*Buchner*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 22 Rn. 11; *Kugelmann*, DuD 2016, 566 (569); *Kumkar*/*Roth-Isigkeit*, JZ 2020, 277 (278); Gola/*Schulz*, DS-GVO, 2. Aufl. (2018), Art. 22 Rn. 3.

[37] Artikel-29-Datenschutzgruppe, WP251rev.01 (2018), S. 21; *Abel*, ZD 2018, 304 (305); HK/DS-GVO/BDSG/*Atzert*, 2. Aufl. (2020), Art. 22 Rn. 2; Kühling/Buchner/*Buchner*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 22 Rn. 12; Paal/Pauly/*Martini*, DS-GVO/BDSG, 3. Aufl. (2021), DS-GVO Art. 22 Rn. 1, 29a f.; Simitis/Hornung/Spiecker gen. Döhmann/*Scholz*, Datenschutzrecht (2019), Art. 22 DSGVO Rn. 16.

[38] Vgl. ABl. EG 1991, Nr. C 159/38 (43). Siehe auch *Bachmeier*, RDV 1995, 49 (51); krit. *Wuermeling*, DB 1996, 663 (668): „sachfremde Regelung".

[39] HK/DS-GVO/BDSG/*Atzert*, 2. Aufl. (2020), Art. 22 Rn. 4 ff.; Kühling/Buchner/*Buchner*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 22 Rn. 11; *Hoeren*/*Niehoff*, RW 2018, 47 (53); *Martini*, JZ 2017, 1017 (1019); Simitis/Hornung/Spiecker gen. Döhmann/*Scholz*, Datenschutzrecht (2019), Art. 22 DSGVO Rn. 3; eingehend *Ernst*, JZ 2017, 1026 (1030 f.); kritisch unter Berufung auf ein nicht hinreichend klar definiertes Schutzziel *Dammann*, ZD 2016, 307 (313).

Entscheidungsprozesse, die dem Betroffenen die Möglichkeit nimmt, den eigenen Standpunkt darzulegen, soll die Letztentscheidungsbefugnis bei einer natürlichen Person liegen.[40] Diese Stoßrichtung verdeutlicht, dass es insoweit weniger um den Schutz der personenbezogenen Daten des Betroffenen und damit um den Schutz des Rechts auf informationelle Selbstbestimmung im Sinne einer Entscheidungsfreiheit darüber geht, welche „eigenen" Daten von Dritten wie verarbeitet werden dürfen. Vielmehr bezweckt Art. 22 Abs. 1 DSGVO den Schutz vor der Unterworfenheit unter eine bestimmte Art der Entscheidungsfindung durch Dritte, ohne indes die Qualität der Eingangsdaten sowie der verwendeten Algorithmen zu regeln.[41]

Art. 22 Abs. 1 DSGVO betrifft zumindest in seinem Kern nur solche Entscheidungen, die ein algorithmisches System vollautomatisch vornimmt, während der Einsatz als Entscheidungsunterstützungsinstrument und damit als bloßes Hilfsmittel im Vorfeld einer menschlichen Entscheidung von dieser Vorschrift nicht erfasst wird.[42] Nun trifft das AMAS selbst keine abschließende Entscheidung über die Zuordnung einer arbeitsuchenden Person zu einer der drei genannten Kategorien, nach der sich wiederum die Fördermaßnahmen des AMS richten. Stattdessen wird die finale Entscheidung erst vom jeweiligen Sachbearbeiter getroffen. Bei einer restriktiven Auslegung kommt Art. 22 Abs. 1 DSGVO daher von vornherein nicht zum Tragen, weil keine Situation vorliegt, bei der es „ohne jegliches menschliche Eingreifen"[43] zu einer Entscheidung kommt. Allerdings besteht weithin Einigkeit darüber, dass eine lediglich am äußeren Wortlaut der Vorschrift haftende Interpretation zu eng ist und diejenigen Konstellationen noch unter diese Bestimmung fallen, bei denen trotz eines formalen menschlichen Letztentscheidungsrechts in materieller Hinsicht eine Entscheidung durch das algorithmische System erfolgt, weil die natürliche Person die Vorgabe des Systems mehr oder weniger mechanisch ohne eigenen Entscheidungsspielraum übernimmt.[44] Damit korrespondierend greift nach Ansicht der Artikel-29-Datenschutzgruppe (nunmehr Europäischer Datenschutzausschuss) Art. 22 DSGVO auch dann ein, wenn ein automatisch erstelltes Profil durch eine Person routinemäßig übernommen wird bzw. die Einbeziehung einer Person nur eine „symbolische Geste" darstellt.[45]

---

[40] Vgl. *Ernst*, JZ 2017, 1026 (1030); siehe auch BT-Drs. 16/10529, S. 13 (zu § 6a BDSG a. F.).

[41] *Broemel/Trute*, Berliner Debatte Initial 27 (2016), 50 (58); *Ernst*, JZ 2017, 1026 (1031).

[42] Siehe nur Paal/Pauly/*Martini*, 3. Aufl. (2021), DS-GVO Art. 22 Rn. 20; Simitis/Hornung/Spiecker gen. Döhmann/*Scholz*, Datenschutzrecht (2019), Art. 22 DSGVO Rn. 28; Gola/*Schulz*, DS-GVO, 2. Aufl. (2018), Art. 22 Rn. 13.

[43] Vgl. ErwGr 71 S. 1 DSGVO.

[44] In diese Richtung – wenn auch im Detail unterschiedlich – die Kommentarliteratur; vgl. HK/DS-GVO/BDSG/*Atzert*, 2. Aufl. (2020), Art. 22 Rn. 75; Kühling/Buchner/*Buchner*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 22 Rn. 15; Paal/Pauly/*Martini*, 3. Aufl. (2021), DS-GVO Art. 22 Rn. 17 ff.; Simitis/Hornung/Spiecker gen. Döhmann/*Scholz*, Datenschutzrecht (2019), Art. 22 DSGVO Rn. 26 f.; Gola/*Schulz*, DS-GVO, 2. Aufl. (2018), Art. 22 Rn. 14; Taeger/Gabel/*Taeger*, DSGVO/BDSG, 3. Aufl. (2019), DSGVO Art. 22 Rn. 26; Däubler/Wedde/Weichert/Sommer/*Weichert*, EU-DSGVO und BDSG, 2. Aufl. (2020), DSGVO Art. 22 Rn. 25.

[45] Artikel-29-Datenschutzgruppe, WP251rev.01 (2018), S. 22.

In diesem Sinne hat die Europäische Kommission schon in ihrem Vorschlag für ein grundsätzliches Verbot automatisierter Entscheidungen in der früheren Datenschutzrichtlinie ausgeführt, dass für die menschliche Beurteilung „Raum bleiben" müsse.[46] An alledem wird deutlich, dass eine Heranziehung von Art. 22 Abs. 1 DSGVO nicht schon deshalb ausscheidet, weil zwischen dem Systemoutput einerseits und der abschließenden Entscheidung andererseits überhaupt eine menschliche Aktivität identifiziert werden kann. Vielmehr muss diese Aktivität so geartet sein, dass der natürlichen Person und nicht dem algorithmischen System die materielle Autorenschaft für die abschließende Entscheidung zukommt. Vor diesem Hintergrund greift es zu kurz, wenn das BVwG lediglich auf die abstrakte Entscheidungsarchitektur abstellt und aus der Existenz interner Richtlinien des AMS für den Umgang der Sachbearbeiter mit dem algorithmisch generierten Entscheidungsvorschlag ohne weiteres ableitet, dass von einer automatisierten Entscheidung im Sinne des Art. 22 Abs. 1 DSGVO nicht die Rede sein kann. Vielmehr gilt es, das soziotechnische System einschließlich der Anwendungspraxis genauer in den Blick zu nehmen, um einschätzen zu können, ob das AMAS entsprechend seiner Intention tatsächlich nur als Hilfsinstrument fungiert oder ob es strukturell und nicht nur im Einzelfall die finale Entscheidung wesentlich mitbestimmt.

Wendet man in diesem Kontext verhaltensökonomische Erkenntnisse an, spricht vieles dafür, dass auch beim AMS-Algorithmus das schon seit langem bekannte „Automation Bias" wirkt, nach dem Menschen dazu neigen, technisch generierten Entscheidungsvorschlägen im Allgemeinen eine Richtigkeitsvermutung zuzusprechen, die sie dazu veranlasst, diesen Vorschlägen auch ohne eine formale Bindung überwiegend zu folgen.[47] Demgemäß sind bei einem in Polen vor einigen Jahren eingeführten (nach einer gerichtlichen Intervention mittlerweile allerdings wieder abgeschafften) vergleichbaren System die Mitarbeiter der dortigen Arbeitsverwaltung nur in einem von rund 200 Fällen von der Systemempfehlung abgewichen, sodass die Abweichungsquote gerade einmal 0,58 % betragen hat,[48] wofür man sich gewiss nicht einfach auf die Erklärung zurückziehen kann, die automatisierten Empfehlungen seien eben „objektiv richtig" gewesen. Verstärkt wird der skizzierte Effekt durch die konkrete Einbettung des AMAS in die praktische Tätigkeit der AMS-Mitarbeiter. So erscheint das Ergebnis der Anwendung des AMS-Algorithmus auf die konkreten Daten des Arbeitsuchenden wie erwähnt gleich zu Beginn der Einzelfallbearbeitung auf der Ein-

---

[46] KOM(92), 422 endg., S. 26.

[47] Vgl. Datenethikkommission, Gutachten (2019), S. 162, 192, 213; ferner *Sommerer*, Personenbezogenes Predictive Policing (2020), S. 71 ff. m. w. N. Zu diesem Phänomen nunmehr auch die Europäische Kommission in Art. 14 Abs. 4 Buchst. a des Vorschlags für eine KI-Verordnung, COM(2021) 206 final, S. 57.

[48] Vgl. *Allhutter* et al. (Fn. 19), S. 89 f. Zum polnischen System eingehend *Niklas*/*Sztandar-Sztanderska*/*Szymielewicz*, Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making (2015).

gabemaske. Diese Gestaltung der Entscheidungssituation begünstigt mithin den sog. Ankereffekt, der das verhaltensökonomische Phänomen bezeichnet, dass bei einem Mangel an klaren und eindeutigen Daten die erste greifbare Größe häufig als Anhaltspunkt für weitere Überlegungen dient.[49] Darüber hinaus ist die Eingabemaske wie geschildert so konfiguriert, dass der Mitarbeiter im Rahmen der Einzelfallbearbeitung nur eine Abweichung von der durch das AMAS vorgeschlagenen Kategorisierung begründen muss, während eine Befolgung der Empfehlung keinen weiteren Begründungsaufwand erfordert. Das System setzt damit in der Sache einen Default-Standard, der die finale Entscheidung bis zu einem gewissen Grad beeinflusst, weil sich der Sachbearbeiter bewusst für ein Opt-out entscheiden muss, wenn er die Systemempfehlung wegen der beim Arbeitsuchenden vorliegenden Besonderheiten überspielen will.[50] Schließlich steht dem jeweiligen Mitarbeiter pro Beratung teilweise nur ein sehr begrenzter Zeitraum zur Verfügung.[51] Alle diese Faktoren bewirken, dass aus der „Zweitmeinung" des algorithmischen Systems letztlich vielfach eine „Erstmeinung" wird.

Ob die Eigenheiten beim Einsatz des AMAS es im Ergebnis rechtfertigen, Art. 22 DSGVO anzuwenden, ist freilich unsicher und hängt neben einer Betrachtung der tatsächlichen Gegebenheiten beim AMS maßgeblich davon ab, ob man nicht nur algorithmendeterminierte Entscheidungssysteme, sondern auch algorithmengetriebene Entscheidungssysteme zumindest unter bestimmten Voraussetzungen bereits nach geltendem Recht unter diese Vorschrift fasst oder aber insoweit lediglich für eine Reform der Regelung plädiert.[52] Zu größerer Klarheit dürfte ein aktuelles Vorlageverfahren des VG Wiesbaden an den EuGH zur sachlichen Reichweite von Art. 22 DSGVO[53] beitragen. Der Anlassfall betrifft die Berechnung der Kreditwürdigkeit einer Person anhand der Anwendung eines automatisierten Verfahrens über die Bildung von Personengruppen auf die konkreten Merkmale des Kreditantragstellers (Scoring).[54] Da die Entscheidung über die Kreditvergabe formal letztlich ebenfalls durch eine natürliche Person erfolgt, ein negativer Score-Wert aber nahezu ausnahmslos zur Versagung eines Kredits führt, das System die Entscheidung in diesen Fällen also faktisch determiniert, handelt es sich um eine in ihren Struktur vergleichbare Situation, sodass mit Spannung erwartet werden darf, ob es der EuGH bei einem formalen Verständnis von Art. 22 DSGVO bewenden lässt oder dem vom VG Wiesbaden präferierten mate-

---

[49] Zum sog. Ankereffekt (Anchoring) siehe etwa *Englerth/Towfigh*, in: Towfigh/Petersen (Hrsg.), Ökonomische Methoden im Recht, 2. Aufl. (2017), Rn. 521 ff.

[50] Vgl. *Allhutter*, WISO 1/2021, 81 (91 ff.); *Allhutter* et al. (Fn. 19), S. 68 ff. Zum verhaltensökonomischen Phänomen des „Status Quo Bias" und der daraus resultierenden Bedeutung von Default Rules näher *Hacker*, Verhaltensökonomik und Normativität (2017), S. 85 f.

[51] Vgl. *Allhutter* et al. (Fn. 19), S. 78 (zehn Minuten).

[52] Siehe hierzu Datenethikkommission, Gutachten (2019), S. 192.

[53] VG Wiesbaden v. 01.10.2021 – 6 K 788/20, ZD 2022, 121. Dazu *Horstmann/Dalmer*, ZD 2022, 260 ff.

[54] Zum Scoring näher *Härting*, IRTB 2016, 209 ff.

riellen Verständnis folgt, was wiederum die Anwendbarkeit von Art. 22 DSGVO auch auf das AMAS wahrscheinlicher machen würde.[55]

Nun würde eine Subsumtion des AMS-Algorithmus unter Art. 22 DSGVO keineswegs zur definitiven Unzulässigkeit dieses Systems führen. Allerdings würden die Rechtmäßigkeitsanforderungen erheblich steigen, weil eine automatisierte Entscheidungsfindung im einschlägigen mitgliedstaatlichen Recht ausdrücklich vorgesehen sein muss, wie sich ErwGr 71 S. 3 DSGVO entnehmen lässt. Weiter müssen diejenigen mitgliedstaatlichen Rechtsvorschriften, die automatisierte Entscheidungen erlauben, angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der Betroffenen enthalten (Art. 22 Abs. 2 Buchst. b DSGVO). Bei der Verarbeitung gesundheitsbezogener Daten müssen entsprechende Maßnahmen zudem nicht nur rechtlich vorgesehen, sondern vom Verantwortlichen als generelle Voraussetzung für die Zulässigkeit einer automatisierten Entscheidung auch tatsächlich getroffen werden (Art. 22 Abs. 4 DSGVO).[56] Dabei würde insbesondere eine Nachschärfung der legislativen Grundlagen für den Einsatz des AMAS dazu führen, die Allokation von Maßnahmen der aktiven Arbeitsmarktpolitik mittels Algorithmen nicht primär aus einer technokratischen Perspektive zu betrachten, sondern in den parlamentarischen Diskurs zu überführen und dadurch die erforderliche Legitimation zu verbreitern.

IV. Weitere Risiken des Einsatzes algorithmischer Systeme

Mit dem Datenschutzrecht wird freilich nur ein Aspekt beim Einsatz des AMAS beleuchtet. Tatsächlich lassen sich noch weitere Risiken und Problemfelder identifizieren, die mit einem algorithmischen System zur Berechnung von Arbeitsmarktchancen und einer darauf aufbauenden Verteilung von Maßnahmen der aktiven Arbeitsmarktpolitik verbunden sind. Dabei steht die Legitimität der eingangs bereits genannten grundsätzlichen Zielbestimmung, die für die aktive Arbeitsmarktpolitik zur Verfügung stehenden personellen und finanziellen Res-

---

[55] Gegen eine großzügige Interpretation von Art. 22 DSGVO lassen sich nunmehr allerdings Art. 6 Abs. 1 Buchst. b und Art. 8 Abs. 1 sowie die ErwGr 32, 35 und 37 des Vorschlags der Europäischen Kommission zur Verbesserung der Arbeitsbedingungen in der Plattformarbeit ins Feld führen, die im Zusammenhang mit automatisierten Entscheidungssystemen ausdrücklich sowohl von *getroffenen* als auch von *unterstützten* Entscheidungen sprechen und hierdurch zu erkennen geben, dass das Unionsrecht durchaus zwischen beiden Kategorien unterscheidet, vgl. COM(2021) 762 final, S. 32 ff., 42 u. 44. Siehe dazu auch *Krause*, NZA 2022, 521 (530).

[56] So Kühling/Buchner/*Buchner*, DS-GVO/BDSG, 3. Aufl. (2020), DS-GVO Art. 22 Rn. 47; Simitis/Hornung/Spiecker gen. Döhmann/*Scholz*, Datenschutzrecht (2019), Art. 22 DSGVO Rn. 64; anders (Regelung in gesetzlicher Erlaubnis ausreichend) HK/DS-GVO/BDSG/*Atzert*, 2. Aufl. (2020), Art. 22 Rn. 157; Paal/Pauly/*Martini*, DS-GVO/BDSG, 3. Aufl. (2021), DS-GVO Art. 22 Rn. 41. Zum Sonderproblem der Verarbeitung von aus „einfachen" und „sensiblen" personenbezogenen Daten bestehenden Mischdatensätzen siehe *Martini*/*Kienle*, JZ 2019, 235 (239).

sourcen effizient und zielgerichtet einzusetzen, außer Frage. Dass eine solche Zielsetzung prinzipiell im öffentlichen Interesse liegt, wird denn auch weder von der Datenschutzbehörde noch von den sonstigen Kritikern des AMAS bezweifelt.

Von zentraler Bedeutung ist indes der Einwand, die vom AMAS erstellten Entscheidungsvorschläge würden zumindest in einer Reihe von Fällen auf Verzerrungen (Bias) beruhen und dadurch unter Umständen sogar zu regelrechten Diskriminierungen führen, also arbeitsuchende Personen infolge von Merkmalen benachteiligen, die nach den allgemeinen Wertungen des Antidiskriminierungsrechts gerade nicht zum Anlass für eine Differenzierung genommen werden dürfen.[57] Die Datenschutzbehörde hatte dieses Problemfeld nicht aufgegriffen, obwohl auch das Datenschutzrecht mit Art. 5 Abs. 1 Buchst. a (Verarbeitung personenbezogener Daten nach Treu und Glauben) sowie mit ErwGr 71 S. 6 a. E. DSGVO (Verhinderung diskriminierender Wirkungen) durchaus Potenzial für eine Verschränkung von Datenschutzrecht und Antidiskriminierungsrecht enthält.[58] Im Einzelnen geraten insoweit insbesondere die in der Modellierungsphase des AMS-Algorithmus verwendeten „verdächtigen" Kriterien Geschlecht, Alter und Behinderung in den Blick. Hinzu drohen mittelbare Benachteiligungen durch die Berücksichtigung weiterer Merkmale. So werden Betreuungspflichten als (negativer) relevanter Parameter nur bei Frauen und damit geschlechtsbezogen berücksichtigt, während der Rekurs auf die Staatengruppe, aus der die arbeitsuchende Person stammt, mittelbar an die ethnische Zugehörigkeit anknüpft. Damit wird ein grundsätzliches Problem algorithmischer Systeme angesprochen, nämlich das ihnen inhärente Risiko diskriminierender Effekte.

Diesbezüglich wird man sich im vorliegenden Zusammenhang nicht mit der Überlegung begnügen können, dass durch die unterschiedlichen Kategorisierungen den davon betroffenen arbeitsuchenden Personen ohnehin nur diejenigen Beratungsleistungen und sonstigen Angebote zugutekommen, die ihrer persönlichen Situation am besten entsprechen, ihnen also kein Nachteil zugefügt wird und die Frage nach einer diskriminierenden Wirkung deshalb gegenstandslos ist. Immerhin stellt es für die Betroffenen einen nicht unerheblichen Unterschied dar, ob sie der mittleren Gruppe zugeordnet werden, auf die sich das Beratungs- und Unterstützungsangebot des AMS konzentrieren soll oder ob sie entweder – infolge einer Zuordnung zur höchsten Gruppe – nicht in den Genuss dieser Leistungen kommen oder – aufgrund einer Zuordnung zur niedrigsten Gruppe – einem gesonderten und offenkundig arbeitsmarktferneren Betreuungsformat zugewiesen werden. Die Auswirkungen der finalen Entscheidung des AMS auf den Arbeitsuchenden sind aufgrund der Bedeutung des für das mittlere Segment vorgesehenen Förderspektrums für das weitere berufliche Fortkommen sowie die Integration in den Arbeitsmarkt somit alles andere als trivial. Vielmehr ist das in Rede stehende

[57] Hierzu eingehend *Allhutter* et al. (Fn. 19), S. 33 ff.

[58] Vgl. *Hacker*, Common Market Law Review 55 (2018), 1143 (1172 f.).

algorithmische System im Sinne der Kritikalitätspyramide der Datenethikkommission zu den Anwendungen mit deutlichem Schädigungspotential[59] bzw. im Sinne der Europäischen Kommission in ihrem Vorschlag für eine Verordnung zur Regulierung Künstlicher Intelligenz zum Hochrisiko-Bereich zu rechnen[60].

Ausgangspunkt für die zunehmend geführte Diskussion über diskriminierende Effekte algorithmischer Entscheidungssysteme[61] ist ihre prinzipielle Funktionsweise. Diese beruht vereinfacht gesagt darauf, dass in der Modellierungsphase eine (möglichst) große Anzahl an Trainingsdaten in das System eingegeben werden, um anhand dieser Daten zu ermitteln, ob im jeweiligen Beobachtungsfeld bestimmte Muster im Sinne statistischer Wahrscheinlichkeiten identifiziert werden können,[62] sodass aus der späteren Zuordnung des konkreten Falls zu einem der ermittelten Muster eine Prognose erstellt werden kann, die wiederum als Grundlage dafür dient, wie in diesem einen Fall entschieden werden soll. Eine solche Vorgehensweise führt zunächst zu einem Phänomen, das als technischer Bias bezeichnet wird und bei dem es darum geht, dass datenbasierte Modellierungen notwendigerweise mit Vereinfachungen der Realität arbeiten müssen, um operabel zu bleiben. So nimmt der AMS-Algorithmus nur eine vergleichsweise grobe Einteilung der Arbeitsuchenden in (lediglich) drei Altersgruppen vor (bis 29 Jahre, 30 bis 49 Jahre und ab 50 Jahre), was im einzelnen Fall zur Fehlklassifikationen von Personen führen kann, wenn etwa einer 49-jährigen Person möglicherweise (deutlich) bessere Arbeitsmarktchancen als einer 50-jährigen Person attestiert werden. Auch können mithilfe eines algorithmischen Systems streng genommen nur Korrelationen zwischen dem Vorhandensein bestimmter Ausgangsdaten einerseits und dem Vorhandensein bestimmter sozialer Zustände und Entwicklungen andererseits ermittelt, nicht aber Kausalitäten dahingehend festgestellt werden, dass der jeweilige Zustand bzw. die jeweilige Entwicklung tatsächlich durch die Existenz bestimmter Ausgangsdaten verursacht worden ist. Ein Sonderproblem besteht zudem darin, dass disruptive Veränderungen auf dem Arbeitsmarkt, etwa der Einbruch im Gastronomiesektor und Hotelleriegewerbe infolge der Corona-Pandemie, bei einer Methodik, die ausschließlich auf in der

---

[59] Datenethikkommission, Gutachten (2019), S. 177 ff.

[60] Vgl. Anhang III Nr. 5 Buchst. a KI-VOE, COM(2021) 206 final, Annexes 1 to 9, S. 4.

[61] Grdl. *Friedman*/*Nissenbaum*, ACM Transactions on Information Systems 14 (1996), 330 ff.; ferner etwa *Barocas*/*Selbst*, California Law Review 104 (2016), 671 ff.; *Geslevich Packin*/*Lev-Aretz*, in: Barfield/Pagallo (Hrsg.), Research Handbook on the Law of Artificial Intelligence (2018), S. 88 (97 ff.); *Hacker*, Common Market Law Review 55 (2018), 1143 ff.; *Hildebrandt*, Smart Technologies and the End(s) of Law (2015), S. 93 ff.; *Kim*, William & Mary Law Review 58 (2017), 857 ff.; *Kollek*/*Orwat*, Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick, TAB-Hintergrundpapier Nr. 24 (2020); *Kroll* et al., University of Pennsylvania Law Review 165 (2017), 633 (678 ff.); *Mayson*, Yale Law Journal 128 (2019), 2218 ff.; *Restrepo Amariles*, in: Barfield (Hrsg.), The Cambridge Handbook of the Law of Algorithm (2021), S. 273 (280 ff.); *Tinhofer*, DRdA 2022, 171 ff.; mit einem Fokus auf die KI-VO-E *Sesing*/*Tschech*, MMR 2022, 24 ff.

[62] Anschaulich *Lopez*, Merkur 863 (2021), 42 (45 f.). Aufschlussreich aus einer gesellschaftstheoretischen Perspektive *Nassehi*, Muster (2019).

Vergangenheit zum Vorschein kommende Muster setzt, nicht zeitnah abgebildet werden können. Nun ist zuzugeben, dass eine feiner ausdifferenzierte Kriterienbildung, die der Komplexität der realen und sich ständig ändernden Verhältnisse auf dem Arbeitsmarkt näherkommen würde, eine Verkleinerung der relevanten Gruppen zur Folge hätte und dadurch wiederum die Basis für die Bildung von Wahrscheinlichkeitsurteilen verringern würde. Immerhin sollte schon das technische Bias Grund genug dafür sein, sich bei der Verwendung solcher Systeme aufgrund ihrer Tendenz zur sachlichen und zeitlichen Abdichtung des Beobachtungsraums[63] der begrenzten Aussagekraft der mithilfe algorithmischer Verfahren gewonnenen Entscheidungsvorschläge bewusst zu sein und und diesem Umstand deshalb durch die Ausgestaltung der Entscheidungsarchitektur hinreichend Rechnung zu tragen.

Noch gravierender ist das Problem des gesellschaftlichen Bias. Indem zur Modellierung des AMS-Algorithmus die arbeitsmarktrelevanten Daten der letzten Jahre verwendet werden, entstehen auf diese Weise Muster, welche die gesellschaftlichen Verwerfungen automatisch widerspiegeln. Dieser Mechanismus ist umso brisanter, als es bei den verwendeten Kriterien nicht lediglich um „neutrale" Merkmale wie Berufsgruppe, Qualifikation und Arbeitsmarktregion, sondern auch um unmittelbar diskriminierungsrelevante Kriterien wie Geschlecht, Alter und Behinderung geht. Wenn aus diesen Daten Wahrscheinlichkeitsurteile im Hinblick auf eine erfolgreiche Integration bzw. Reintegration in den Arbeitsmarkt gewonnen werden und sich daran die Verteilung von Ressourcen der aktiven Arbeitsmarktpolitik ausrichtet, besteht ein erhebliches Risiko, dass sich strukturelle Benachteiligungen auf dem Arbeitsmarkt fortsetzen und verfestigen.[64] Denn ein System, das der Aufdeckung existierender sozioökonomischer Muster dient und dass sich deshalb definitionsgemäß an der Vergangenheit orientiert, läuft automatisch Gefahr, vorhandene soziale Ungleichheiten in der Zukunft fortzuschreiben und Lebenslagen ohne Rücksicht darauf zu reproduzieren, ob sie durch diskriminierende Verhaltensweisen entstanden sind.[65]

Nun sollen algorithmische Systeme zunächst einmal genau dies leisten, nämlich die realen Verhältnisse bestmöglich abbilden, was bei einem Mechanismus wie dem AMAS zwangsläufig dazu führt, dass strukturelle Schlechterstellungen unterschiedlicher Gruppen von Arbeitsuchenden in die Musterermittlung einfließen. Plakativ formuliert ist es danach nicht der AMS-Algorithmus selbst, der diskriminiert. Vielmehr fördere ein solches System nur vorhandene gesellschaftliche Diskriminierungen zutage. Gegen eine solche Argumentation ist freilich einzuwenden, dass die Vorstellung, Daten über soziale Zustände würden die Realität

---

[63] *Büchner/Dosdall*, Köln Z Soziol 73 (2021), Suppl 1, 333 (340).

[64] *Allhutter*, WISO 1/2021, 81 (89 ff.).

[65] Vgl. *Büchner/Dosdall*, Köln Z Soziol 73 (2021), Suppl 1, 333 (340). Dazu anschaulich auch *O'Neil*, Weapons of math destruction (2016), S. 204: „Big Data processes codify the past. They do not invent the future".

lediglich widerspiegeln, nicht aber ihrerseits die Realität strukturieren, die Sachlage zu sehr vereinfacht.[66] Zum einen ist schon die Auswahl der für die Modellierung verwendeten Trainingsdaten nicht mit der schlichten Messung naturwissenschaftlich relevanter Phänomene vergleichbar.[67] Vielmehr ist die Selektion der für die Bildung eines algorithmischen Systems herangezogenen Daten auf dem Gebiet sozialer Prozesse eine ausgesprochen voraussetzungsvolle Angelegenheit. So ist etwa die Frage, ob überhaupt eine Differenzierung nach dem Geschlecht stattfinden soll, keine „neutrale" Vorentscheidung, sondern eine bewusste Setzung seitens des AMS sowie der Systementwickler, die unterschiedlich ausfallen kann, jedenfalls aber bewusst und nicht aufgrund einer unreflektierten Binäreinteilung der arbeitsuchenden Personen erfolgen sollte. Anders gesagt können sich schon in der Phase der Datenaggregation diskriminierende Effekte einschleichen, die sich auf die anschließende Modellierung des algorithmischen Systems auswirken können und die den ausgeworfenen Ergebnissen ein entsprechendes Bias verleihen. Zum anderen ist die potenziell diskriminierende Wirkung des AMAS kein bloßes Beiprodukt, sondern entspricht der inneren Logik des Systems, baut die gewählte Methode doch im Kern auf der Grundannahme auf, dass sich der für die Vergangenheit angenommene Zusammenhang zwischen bestimmten gruppenbezogenen Merkmalen und schlechteren Integrationschancen auch in Zukunft bei denjenigen Personen fortsetzen wird, die aufgrund der ihnen zugeschriebenen Merkmale der betreffenden Gruppe zuzuordnen sind. Eine derartige Prognose muss aber keineswegs in allen Einzelfällen zutreffen. Bei einer von der konkreten Situation abstrahierenden Betrachtungsweise besteht somit die Gefahr eines „Generalisierungsunrechts"[68]. Darüber hinaus riskieren algorithmenbasierte Chancenberechnungen zu einer self-fulfilling prophecy zu werden, indem sie für die eine Gruppe Türen öffnen und für die andere Gruppe Türen verschließen[69] sowie dadurch demotivierende Wirkung entfalten können, dass einem Arbeitsuchenden mit dem Nimbus der Objektivität bescheinigt wird, aufgrund individueller Merkmale[70] zu einer Gruppe von Menschen zu gehören, die auf dem Arbeitsmarkt von vornherein nur geringe Chancen haben[71]. Nicht zuletzt führt eine strikte Kategorien-

---

[66] *Hagendorff/Wezel*, AI & Society 35 (2020), 355 (356). Zur impliziten Normativität empirischer Aussagen allgemein *Hamann*, Evidenzbasierte Jurisprudenz (2014), S. 107 ff.; zur Empirie als soziales Konstrukt ferner eindringlich *Augsberg*, Der Staat 51 (2012), 117 ff.; umfassend *Luhmann*, Erkenntnis als Konstruktion (1988).

[67] Zur dahinterstehenden – zumeist unreflektierten – Annahme, soziale Zustände ähnlich wie physikalische Zustände messen zu können, anschaulich *Desrosières*, Social Research 68 (2001), 339 (340 ff.) („metrological realism"); hierzu auch *Hildebrandt* (Fn. 61), S. 36; *Lorentz* (Fn. 31), S. 71.

[68] Vgl. *Britz*, Einzelfallgerechtigkeit versus Generalisierung (2008), S. 2.

[69] Zu diesem Effekt *Hacker*, Common Market Law Review 55 (2018), 1143 (1150); *Kim*, William & Mary Law Review 58 (2017), 857 (894).

[70] Die Individualisierung struktureller Diskriminierungen als Folge algorithmenbasierter Entscheidungen betonen *Büchner/Dosdall*, Köln Z Soziol 73 (2021), Suppl 1, 333 (342).

[71] *Allhutter* et al. (Fn. 19), S. 77 f.

bildung unweigerlich zu einer hierarchischen Vorstellung von „besseren" und „schlechteren" Gruppen von Arbeitsuchenden und leistet damit einer sozialen Stigmatisierung Vorschub.[72]

Obgleich die Risiken somit nicht zu unterschätzen sind, sollte man umgekehrt das emanzipatorische Potenzial von algorithmischen Systemen zur Ermittlung der Arbeitsmarktchancen verschiedener Personengruppen nicht ausblenden. So können solche Systeme dazu beitragen, vermutete Muster von Diskriminierungen auf dem Arbeitsmarkt statistisch zu belegen oder sogar bislang unentdeckte Diskriminierungsmuster erstmals freizulegen, um auf dieser Basis Gegenstrategien zur Beseitigung sozial unerwünschter Benachteiligungen zu entwickeln. Je genauer die Kenntnis darüber ist, welche Unterschiede in der Vergangenheit am Arbeitsmarkt eine Rolle bei der gesellschaftlichen Zuteilung von vorteilhaften Positionen gespielt haben, desto genauer lässt sich bestimmen, welche Unterschiede in der Zukunft möglichst keine Bedeutung mehr haben sollen und welcher Maßnahmen es bedarf, um dieses sozialpolitische Ziel zu erreichen. Dieser Aspekt zeigt im Übrigen einmal mehr, dass sich eine Bewertung nicht auf den AMS-Algorithmus als solchen beschränken sollte, sondern dass auch die Einbettung in die sonstige Entscheidungsarchitektur und damit das gesamte soziotechnische System in den Blick zu nehmen ist.

V. Lösungsansätze

Lässt man die skizzierten Problemfelder Revue passieren, scheint sich zunächst das Datenschutzrecht als Lösungsansatz anzubieten. Hierbei geht es vor allem um die Einbeziehung in den sachlichen Schutzbereich von Art. 22 DSGVO. Ob sich der EuGH aufgrund der erwähnten Vorlage des VG Wiesbaden dazu bewegen lassen wird, die Vorschrift auch auf solche Fälle anzuwenden, in denen eine Entscheidung zwar nicht im wortwörtlichen Sinne ausschließlich auf einer automatisierten Verarbeitung beruht, sondern in denen die fragliche Entscheidung formal durch einen Menschen getroffen wird, diese Entscheidung aber algorithmengetrieben ist und vom System abweichende finale Entscheidungen nur selten vorkommen, ist gegenwärtig allerdings noch offen. Eine insoweit großzügige Interpretation von Art. 22 DSGVO durch den EuGH würde nicht nur eine Nachschärfung der mitgliedstaatlichen Rechtsgrundlagen nach Maßgabe von Art. 22 Abs. 2 Buchst. b und Abs. 4 DSGVO erfordern, sondern auch weitere datenschutzrechtliche Schutzmechanismen mobilisieren. Dies betrifft gemäß Art. 13 Abs. 2 Buchst. b bzw. Art. 14 Abs. 2 Buchst. g DSGVO gesteigerte ex-ante-Informationspflichten des Verantwortlichen gegenüber der betroffenen Person im Zusammenhang mit der Erhe-

[72] Dies heben *Niklas/Sztandar-Sztanderska/Szymielewicz* (Fn. 48), S. 36 f., für das polnische System hervor.

bung von personenbezogenen Daten, die sich insbesondere auf die „involvierte Logik" der automatisierten Entscheidungsfindung beziehen.[73] Hinzu tritt nach Art. 15 Abs. 1 Buchst. h DSGVO ein entsprechend gesteigertes ex-post-Auskunftsrecht des Betroffenen.[74] Dagegen gehören die in Art. 22 Abs. 3 DSGVO genannten Rechte, nämlich das Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen sowie auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung nicht automatisch zum Schutzprogramm, weil sich diese Regelung nur auf Art. 22 Abs. 2 Buchst. a und c DSGVO, nicht aber auf Art. 22 Abs. 2 Buchst. b DSGVO bezieht. Zudem wäre ein Recht auf eine menschliche Intervention („Human in the Loop") im Kontext des AMAS praktisch weithin gegenstandslos, weil die finale Entscheidung über die jeweils zu ergreifenden arbeitsmarktpolitischen Maßnahmen nach den internen Vorgaben des AMS formal schon jetzt beim jeweiligen Sachbearbeiter liegt. Auch stand ein Ausschluss von Widerspruchsmöglichkeiten des Betroffenen ohnehin nie zur Debatte. Allerdings ist bereits grundsätzlich in Zweifel zu ziehen, ob ein Schutz, der auf eine Stärkung der Rechte des einzelnen Betroffenen setzt und ein höheres Maß an individueller Transparenz, Erkennbarkeit und Nachvollziehbarkeit der jeweiligen Entscheidung fordert, tatsächlich ein zielführender Ansatz ist, weil er die Wahrung der eigenen schutzwürdigen Interessen zu großen Teilen der Eigeninitiative des Einzelnen überlässt. Gerade den Angehörigen vulnerabler und marginalisierter Gruppen wird es angesichts ihrer sozioökonomischen Lage sowie den damit verbundenen Problemen und Hemmnissen indes schwerfallen, gegenüber der öffentlichen Arbeitsverwaltung ihre Interessen eigenständig zu artikulieren, wenn sie sich durch den Einsatz eines algorithmischen Systems benachteiligt fühlen.

Die Fokussierung auf den individuellen Rechtsschutz ist auch einer der Gründe, warum das Antidiskriminierungsrecht letztlich nur eingeschränkt geeignet ist, bei etwaigen Benachteiligungen aufgrund von unzulässigen Diskriminierungsmerkmalen infolge des Einsatzes eines algorithmischen Systems für eine effektive Abhilfe zu sorgen. Gegen eine Diskriminierung lässt sich im Ausgangspunkt nämlich nur dann vorgehen, wenn eine benachteiligende Entscheidung darauf zurückgeführt werden kann, dass ein bestimmtes Merkmal im Entscheidungsprozess objektiv eine Rolle gespielt hat. Insoweit genügt zwar nach allgemeinen Grundsätzen eine Mitursächlichkeit.[75] Auch fordert das europäische Antidiskriminierungsrecht innerhalb seines Anwendungsbereichs Beweiserleichterungen,[76] die sich nutzbar machen bzw. ausdehnen ließen. Selbst unter erleichterten Voraus-

---

[73] Zur (umstrittenen) Reichweite der Regelung weiterführend *Wachter/Mittelstadt/Floridi*, International Data Privacy Law 7 (2017), 76 ff.

[74] Aus ErwGr 71 S. 4 DSGVO dürfte indes kein eigenständiges „Recht auf Erläuterung" zu gewinnen sein; vgl. dazu näher *Kumkar/Roth-Isigkeit*, JZ 2020, 277 (280 f.).

[75] Siehe etwa Preis/Sagan/*Grünberger/Husemann*, EUArbR, 2. Aufl. (2019), Rn. 5.127 ff. m. w. N.

[76] Vgl. Art. 8 RL 2000/43/EG, Art. 10 RL 2000/78/EG, Art. 9 RL 2004/113/EG und Art. 19 RL 2006/54/EG.

setzungen wird es indes nicht selten schwierig sein, bei einem algorithmischen System, das aufgrund von im Einzelnen kaum einsehbaren und nachvollziehbaren Rechenoperationen (Black Box-Effekt) lediglich ein bestimmtes Ergebnis aufwirft, hinreichende Anhaltspunkte dafür zu finden, dass eine Benachteiligung gerade auf der unzulässigen Verwendung eines Merkmals der betroffenen Person beruht.

Erkennt man, dass der besondere Charakter des Einsatzes von algorithmischen Systemen zur Berechnung von Arbeitsmarktchancen darin besteht, dass aus vergangenen Umständen durch Pauschalisierung Muster gewonnen und der künftigen Förderpraxis schematisch zu Grunde gelegt werden („Herrschaft der Vergangenheit über die Zukunft"),[77] wird deutlich, dass der Ausbau individueller Rechtspositionen letztlich nur begrenzt weiterführt. Erfolgversprechender erscheint es zum einen, die Anwendung solcher Systeme an objektive Anforderungen zu binden, wie sie nunmehr etwa im Vorschlag der Europäischen Kommission für eine Verordnung zur Regulierung Künstlicher Intelligenz für Hochrisiko-KI-Systeme formuliert werden und zu denen ein effektives Risikomanagementsystem, eine geeignete Daten-Governance und eine hinreichend qualifizierte menschliche Aufsicht gehören.[78] Zum anderen bietet sich gerade beim Einsatz von Instrumenten der öffentlichen Arbeitsverwaltung, die sich massiv auf die Arbeitsmarktchancen zahlreicher Arbeitsuchender auswirken, die Schaffung einer spezifischen Partizipations- und Kontrollarchitektur an. So sollte schon bei der Auswahl des Datenmaterials wie auch bei der anschließenden Frage nach der Relevanz des systemgenerierten Vorschlags in der konkreten Entscheidungssituation auch die Perspektive der Betroffenen einbezogen werden. Dies ließe sich durch eine Einbindung von Interessenvertretungen als Repräsentanten der betroffenen Personen bereits im Entwicklungsstadium solcher Systeme bewerkstelligen, um für Fairness und Transparenz[79] sowie nicht zuletzt für ein hohes Maß an sozialer „Legitimation durch Verfahren" zu sorgen. Weiter könnte eine kritische Begleitung der Entwicklung und Einführung eines für die Arbeitsverwaltung derart zentralen Systems durch neutrale Experten dazu beitragen, dass Anregungen und Einwände sowohl im Hinblick auf die Qualität der verwendeten Eingangsdaten und die Modellierung des Algorithmus mit dem Ziel der Vermeidung von ungerechtfertigten Diskriminierungen („algorithm fairness approaches", „bias detection and corrections strategies", „equal treatment by design")[80] als auch hinsichtlich der

---

[77] Treffend *Lopez*, Merkur 863 (2021), 42 (46 ff.).

[78] Vgl. Art. 8 ff. KI-VOE, COM(2021) 206 final, S. 52 ff. Siehe auch Datenethikkommission, Gutachten (2019), S. 212 ff.

[79] Zur Bedeutung und den verschiedenen Facetten von Fairness und Transparenz beim Einsatz algorithmischer Entscheidungssysteme *Lepri* et al., Philosophy & Technology 31 (2018), 611 (615 ff.). Zu den unterschiedlichen Formen von Opazität siehe auch *Burrell*, Big Data & Society 3 (1) (2016), 1 (4 ff.).

[80] Dazu *Hacker*, Common Market Law Review 55 (2018), 1143 (1175 ff.). Zu Methoden bei der Analyse von Diskriminierungsphänomenen eingehend *Romei*/*Ruggieri*, The Knowledge Engineering Review 29 (2014), 582 ff.

Berücksichtigung verhaltenswissenschaftlicher Erkenntnisse bei der Gestaltung der Entscheidungssituation für den jeweiligen AMS-Mitarbeiter frühzeitig und sachkundig vorgetragen und verhandelt werden. Schließlich empfiehlt es sich, die Erfahrungen beim Einsatz des AMAS nach einer gewissen Zeit wiederum unter Heranziehung neutraler Experten zu evaluieren und bei der Strukturierung des Systems gegebenenfalls nachzusteuern. Diese Maßnahmen dürften nicht nur der Qualität des algorithmischen Systems zugutekommen und damit zu im Ergebnis „besseren" Entscheidungen[81] führen, sondern nicht zuletzt auch dessen gesellschaftliche Akzeptanz deutlich steigern.[82]

VI. Resümee

Algorithmische Entscheidungssysteme auf dem Gebiet der öffentlichen Arbeitsverwaltung wie das österreichische AMAS stehen an der Schnittstelle unterschiedlicher Rechtsgebiete und sozialpolitischer Zielvorstellungen. Auch wenn die durch sie aufgeworfenen Rechtsfragen einer Antwort bedürfen, ist es mit einer rein juristischen Betrachtung nicht getan. In Anlehnung an die Worte der Vizepräsidentin der Europäischen Kommission *Margrethe Vestager*, „On Artificial Intelligence, trust is a must, not a nice to have"[83], ist auf einem derart sensiblen Feld wie der Allokation von Maßnahmen der aktiven Arbeitsmarktpolitik mithilfe eines Algorithmus vielmehr ein hohes Maß an Vertrauen bei allen Beteiligten vonnöten, um die Effizienzpotenziale beim Einsatz eines solchen Instruments umfassend zu heben. Arbeitsmarktchancen sind letztlich Lebenschancen. Dies sollte Anlass genug sein, durch die institutionellen Anforderungen an die Schaffung eines algorithmischen Systems zur Berechnung von Arbeitsmarktchancen und darauf fußenden Fördermaßnahmen sowie insbesondere durch eine entsprechende Partizipations- und Kontrollarchitektur bei den Betroffenen das Vertrauen zu wecken und zu stärken, nicht lediglich anonymen Algorithmen ausgeliefert zu sein, sondern bei den notwendigen Entscheidungen eine individuelle und diskriminierungsfreie Chance zu haben.

---

[81] Vgl. *Kleinberg* et al., Quarterly Journal of Economics 133 (2018), 237 ff.
[82] Weitere Empfehlungen bei *Allhutter* et al. (Fn. 19), S. 99 ff.
[83] European Commission, Europe fit for the Digital Age, Pressemitteilung v. 21.4.2021.

„Transparency by Design" als Rechtsprinzip gegen Dark Patterns

Anne Lauber-Rönsberg[1]

I. Einleitung

In der letzten Zeit ist mehr und mehr in den Fokus der öffentlichen Diskussion gerückt,[2] dass bei der Gestaltung digitaler Nutzeroberflächen z. T. Designmuster verwendet werden, die Verbraucherinnen und Verbraucher mittels kognitiver Heuristiken zu Entscheidungen bewegen, die sie ansonsten nicht getroffen hätten und die ggf. sogar ihren Interessen zuwiderlaufen.[3] Beispiele für solche sog. Dark Patterns oder manipulativen Designs sind das Verschleiern von relevanten Informationen, z. B. durch entsprechende graphische Gestaltungen, Ausgestaltung, bei denen die Wahl der vom Anbieter nicht erwünschten Option aufwändiger ist als die Wahl der erwünschten Option, z. B. die datenschutzrechtliche Einwilligung im Rahmen von Cookie-Bannern, missverständliche Fragestellungen im Rahmen des Vertragsschlusses, die Verführung durch sachfremde Anreize wie ein emotional manipulatives Framing oder die Betonung einer vermeintlichen oder tatsächlich bestehenden Knappheit oder Befristung eines Angebots.[4] Wie diese Beispiele zeigen, kann sich die manipulative Wirkung des Designs entweder aus der grafischen Ausgestaltung der Benutzeroberfläche, aus der Menüführung oder aus den mitgeteilten Inhalten ergeben.

---

[1] Dr. Sven Hetmank und Ass. iur. Franz Lehr danke ich herzlich für hilfreiche Anregungen und Kommentare.

[2] S. z. B. den Bericht des norwegischen Forbrukerradet, Deceived by Design, 2018, https://www.dwt.com/-/media/files/blogs/privacy-and-security-blog/2020/12/deceived-by-design.pdf sowie den Workshop der US Federal Trade Commission am 29.4.2021, https://www.ftc.gov/news-events/events/2021/04/bringing-dark-patterns-light-ftc-workshop.

[3] S. den Definitionsvorschlag für Dark Patterns in der Position des Europäischen Parlaments zum Gesetz über Digitale Dienste vom 20.1.2022 in Erw. 39a, P9_TA(2022)0014.

[4] S. die umfassenden Darstellungen bei *Luguri/Strahilevitz*, Shining a Light on Dark Patterns, JLA 13 (2021), 43 (53); *Mathur/Acar/Friedman/u. a.*, Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites, Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 81 (November 2019); *Martini/Kramme/Seeliger*, „Nur noch für 30 Minuten verfügbar" – Scarcity- und Countdown-Patterns bei Online-Geschäften auf dem Prüfstand des Rechts, VuR 2022, 123; *Martini/Drews/Seeliger/Weinzierl*, Dark Patterns, ZfDR 2021, 47 (52) und die Internetseite des von den Autoren durchgeführten Forschungsprojekts „Dark Pattern Detection Project" (dapde): https://dapde.de/de/dark-patterns/arten-und-beispiele/ sowie die Seite des Dark Pattern-„Entdeckers" *Harry Brignull*, https://www.deceptive.design/.

Das Konzept von manipulativen Designs beruht darauf, dass die Nutzerinnen und Nutzer zwar die relevanten und erforderlichen Informationen erhalten, allerdings durch die Aufbereitung der Informationen und/oder die Ausgestaltung der Entscheidungsprozesse in ihrer Willensbildung beeinflusst werden. Typische Einsatzszenarien sind Internetseiten, Apps, Social Media Accounts[5] sowie Computerspiele, wo solche Ausgestaltungen z. B. auf eine Spielzeitverlängerung, die Erhöhung von Downloads und In-App-Käufe zielen. Ermöglicht wird die Beeinflussung durch die – den Verbrauchern zum Teil nicht bewusste – Asymmetrie der Gestaltungsmacht über die Ausgestaltung von Entscheidungsarchitekturen und Nutzeroberflächen.

Für unterschwelligen Beeinflussungen im analogen wie digitalen Raum haben der Ökonom Thaler und der Rechtswissenschaftler Sunstein den Begriff des Nudging geprägt. Allerdings sollen Nudges nach ihrem Konzept, das einem liberalen Paternalismus verpflichtetet ist, primär dazu eingesetzt werden, um Individuen zu einem „besseren, gesünderen und längeren" Leben zu verhelfen.[6] Werden solche Mechanismen wie in den hier analysierten Beispielen dazu genutzt, um Nutzer und Nutzerinnen zu von ihnen nicht intendierten oder interessenwidrigen Dispositionen zu verleiten, z. B. zu Vertragsabschlüssen oder zur Preisgabe persönlicher Informationen, dann werden diese unterschwelligen Beeinflussungen nach dem Webdesigner Harry Brignull als „Dark Patterns" oder auch als „manipulative Designs" bezeichnet.

Manipulative Designs können z. B. dazu eingesetzt werden, um einen Vertragsschluss zu fördern bzw. die Beendigung eines Vertragsverhältnisses, z. B. durch Kündigung, zu erschweren, um Verbraucher dazu zu bewegen, in die Verarbeitung ihrer personenbezogenen Daten einzuwilligen, eine zuvor erteilte Einwilligung nicht zu widerrufen oder, auf tatsächlicher Ebene, personenbezogene Daten preiszugeben, sowie um Aufmerksamkeit auf bestimmte Inhalte zu lenken. Manipulative Designs können somit zu wirtschaftlichen Nachteilen für Verbraucher führen, wenn es zu einem für sie nicht interessengerechten Vertragsschluss kommt oder eine unangemessen hohe Gegenleistung vereinbart wird. Zum zweiten können manipulative Designs die informationelle Selbstbestimmung beeinträchtigen, indem sie die Preisgabe bzw. die Einwilligung in die Verarbeitung von personenbezogenen Daten fördern. Darüber hinaus können sie zum dritten allgemein zu Einbußen an Autonomie und Entscheidungsfreiheit i. S. der Fähigkeit, eine informierte Entscheidung zu treffen, führen. Insbesondere der letztgenannte Gesichtspunkt zeigt aber auch, dass eine Herausforderung darin besteht, dass nicht jede Form der Beeinflussung von Verbraucherinnen und Verbrauchern gleichermaßen regulierungsbedürftig ist, sondern dass hier eine Grenze zu definieren ist zwischen Beeinflussungsversuchen, die einem autonom und selbstbestimmt agierenden Ver-

[5] S. dazu *EDSA*, Leitlinien 03/2022 zu Dark Patterns in social media platform interfaces, 14.3.2022, Fassung für die öffentliche Konsultation.

[6] *Sunstein/Thaler*, Nudge: Wie man kluge Entscheidungen anstößt, 2009.

braucher – auch angesichts der grundrechtlich geschützten unternehmerischen Freiheit des Verwenders[7] – zuzumuten sind, und Manipulationsversuchen, die eine regulierungsbedürftige Beeinträchtigung der Entscheidungsfreiheit darstellen.

Bislang liegen wenig öffentlich zugängliche Studienergebnisse über die Wirksamkeit von manipulativen Designs von digitalen Oberflächen vor, was wenig überraschend ist, da „reichweitenstarke Online-Anbieter" entsprechende Tests zwar durch Anpassungen ihrer Oberflächengestaltung einfach vornehmen können,[8] entsprechende Studien für Forscherinnen und Forscher, deren normales Alltagsgeschäft nicht in der Bereitstellung von Online-Angeboten besteht, sehr aufwändig sind.[9] Die von Luguri und Strahilevitz durchgeführte Studie belegt aber zum einen die Wirksamkeit von manipulativen Designs zur Steigerung der Anzahl von Vertragsabschlüssen[10] und zum anderen eine größere Anfälligkeit von Nutzern mit einem geringeren Bildungsgrad für manipulative Designs.[11] Es ist zu erwarten, dass es zukünftig durch den Einsatz von KI-Systemen möglich sein wird, Nutzeroberflächen personalisiert an den jeweiligen Nutzer anzupassen und hierdurch die Wirksamkeit manipulativer Designs zu erhöhen.[12] Eine Zukunftsvision, wie die Ansprache von Kunden durch virtuelle Assistenten oder Chatbots künftig ausgestaltet werden könnte, zeigt eine zugunsten von Amazon patentierte Erfindung zur Bestimmung des physischen oder emotionalen Zustands des Nutzers durch die Auswertung seiner „Gespräche" mit einem virtuellen Assistenten sowie weiterer Daten, um diesem dann zielgerichtet Produkte – z. B. Hustenbonbons im Falle einer Erkältung – empfehlen zu können.[13]

Manipulative Designs sind allerdings kein neues, auf digitale Nutzeroberflächen beschränktes Phänomen, sondern kommen ebenso an der Supermarktkasse oder in anderen Szenarien von Vertragsverhandlungen zum Einsatz – Stichwörter sind Überrumpelungssituationen, Ausübung von (moralischem) Druck oder irreführende Präsentation von Informationen. Insofern findet sich im geltenden Lauterkeits- und Datenschutzrecht bereits eine Reihe von Transparenzvorgaben.[14]

---

[7] *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 2022, 22 f., https://www.bevh.org/fileadmin/content/04_politik/Europa/Kuehling-Gutachten-BEVH-Dark-pattern-22-02-16-final.pdf.

[8] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (50).

[9] *Weinzierl*, Dark Patterns als Herausforderung für das Recht, NVwZ-Extra 2020, 1 (3).

[10] Nach den Studienergebnissen abonnierten bei einem neutralen Design nur 11 % der Probanden einen kostenpflichtigen Service, bei einer milden Kombination von Dark Patterns 25 % und bei einer aggressiven 37 %, s. *Luguri/Strahilevitz*, JLA 13 (2021), 43 (65).

[11] *Luguri/Strahilevitz*, JLA 13 (2021), 43 (80 f).

[12] *Luguri/Strahilevitz*, JLA 13 (2021), 43 (103); *Weinzierl*, NVwZ-Extra 2020, 1 (3).

[13] United States Patent No. US 10,096,319 B1 vom 9.10.2018: Voice-based Determination of Physical and Emotional Characteristics of Users.

[14] S. zu den hier nicht behandelten vertragsrechtlichen Ansatzpunkten *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", https://www.bevh.org/fileadmin/content/04_politik/Europa/Kuehling-Gutachten-BEVH-Dark-pattern-22-02-16-final.pdf28 ff.; *Martini/Kramme/Seeliger*, VuR 2022, 123 (127 ff.).

II. Lauterkeitsrecht

Dies gilt an erster Stelle für das Lauterkeitsrecht, das nach der in § 1 UWG festgeschriebenen Schutzzwecktrias neben dem Schutz der Mitbewerber und sonstiger Marktteilnehmer auch dem Schutz der Verbraucherinnen und Verbraucher vor unlauteren geschäftlichen Handlungen dient. So stellen insbesondere Transparenzpflichten in Form von Irreführungsverboten bereits bislang ein zentrales Instrument des lauterkeitsrechtlichen Verbraucherschutzes dar und werden nun durch die zum 28.5.2022 in Kraft tretenden Änderungen durch das Gesetz zur Stärkung des Verbraucherschutzes im Wettbewerbs- und Gewerberecht,[15] das der Umsetzung der RL 2019/2161 zur Modernisierung des Verbraucherschutzrechts dient, nochmals erweitert. Allerdings lassen sich dem Lauterkeitsrecht keine allgemeinen Aussagen über die Zulässigkeit manipulativer Designs entnehmen. Vielmehr ist jedes Pattern für sich an den lauterkeitsrechtlichen Verbotstatbeständen zu messen.

1. Schutzzwecke des Lauterkeitsrechts

Ein interessanter Unterschied zwischen dem UWG und der RL über unlautere Geschäftspraktiken 2005/29, die zu einer Harmonisierung der Regelungen im B2C-Verhältnis geführt hat, ist allerdings, dass sich der Schutzzweck der UGP-Richtlinie auf den Schutz der wirtschaftlichen Interessen der Verbraucher beschränkt (Art. 1 UGP-RL und Erwägungsgründe 6 und 8 UGP-RL).[16] Demgegenüber sind die Schutzzwecke des UWG weiter. Dieses schützt zum einen die geschäftliche Entscheidungsfreiheit der Verbraucher,[17] aber auch sonstige Rechte und Rechtsgüter, z. B. das allgemeine Persönlichkeitsrecht oder das Interesse, dass belästigende Geschäftspraktiken die Aufmerksamkeit oder die Ressourcen des Verbrauchers nicht in unzumutbarer Weise beanspruchen.[18]

Dies wirft die Frage auf, ob das europäische Verbraucherschutzrecht auch für solche Fallkonstellationen, in denen manipulative Designs nicht primär wirtschaftliche Verbraucherinteressen beeinträchtigen, sondern z. B. dazu eingesetzt werden, um den Verbraucher zu einer datenschutzrechtlichen Einwilligung oder zu der faktischen Preisgabe personenbezogener Daten zu motivieren, harmonisiert wurde. Relevant wird dies für die lauterkeitsrechtlichen Tatbestände, die daran anknüpfen, dass der Verbraucher zu einer geschäftlichen Entscheidung ver-

---

[15] Gesetz zur Stärkung des Verbraucherschutzes im Wettbewerbs- und Gewerberecht v. 10.8. 2021, BGBl. I S. 3504.

[16] Köhler/Bornkamm/Feddersen/*Köhler*, Gesetz gegen den unlauteren Wettbewerb, 40. Aufl. 2022, UWG § 1 Rn. 15; MüKoUWG/*Micklitz*/*Namyslowska*, Münchener Kommentar zum Lauterkeitsrecht, 3. Aufl. 2020, UGP-Richtlinie Art. 1 Rn. 1.

[17] Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 1 Rn. 18 f.

[18] Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 1 Rn. 20.

anlasst wird (§ 4a Abs. 1, § 5 Abs. 1, § 5a Abs. 2 UWG) oder die wie § 3 Abs. 2 UWG darauf abstellen, ob eine geschäftliche Handlung dazu geeignet ist, das wirtschaftliche Verhalten des Verbrauchers wesentlich zu beeinflussen.

Unter einer geschäftlichen Entscheidung ist gemäß § 2 Abs. 1 Nr. 9 UWG/Art. 2 lit. k) UGP-RL jede Entscheidung eines Verbrauchers zu verstehen, ob, wie und unter welchen Bedingungen er ein Geschäft abschließen, eine Zahlung leisten, eine Ware oder Dienstleistung behalten oder abgeben oder ein vertragliches Recht im Zusammenhang mit einer Ware oder Dienstleistung ausüben will. Der EuGH fasst hierunter auch mit einem Vertrag unmittelbar zusammenhängende Entscheidungen, die dem Vertragsschluss vorgelagert sind, z. B. zum Betreten eines Geschäfts.[19] Entsprechendes gilt für das Aufsuchen einer Internetseite, auf der Produkte oder Dienstleistungen unmittelbar bestellt werden können, oder eines den Absatz fördernden Instagram-Kanals.[20] Anders entschied der BGH hingegen für den Fall, dass ein Verbraucher aufgrund eines irreführenden ersten Eindrucks, der bei näherem Hinsehen berichtigt wurde, dazu gebracht wird, sich mit einem beworbenen Angebot näher zu befassen.[21] Wenn ein Verbraucher durch eine manipulative, aber sogleich berichtigte Aufmachung dazu gebracht wird, seine knappen Aufmerksamkeitsressourcen auf bestimmte Informationen zu lenken, wäre dies mangels geschäftlicher Entscheidung im Rahmen der genannten Unlauterkeitstatbestände somit nicht relevant – es sei denn, durch das Anlocken würden dem Unternehmer Wettbewerbsvorteile entstehen,[22] wie z. B. wohl die Möglichkeit, Besucher einer Internetseite zu tracken, um zukünftig personalisierte Werbung auszuspielen.

Auch die Entscheidung über die Preisgabe personenbezogener Daten stellt zumindest dann eine geschäftliche Entscheidung dar, wenn die Daten von dem Anbieter zum Zweck der Werbung für eine entgeltliche Dienstleistung erhoben werden.[23] Streitig ist allerdings, ob Entscheidungen des Verbrauchers über die Entgegennahme unentgeltlicher Waren oder Dienstleistungen, z. B. unentgeltliche Online-Angebote oder Bewertungsportale, als geschäftliche Entscheidungen anzusehen sind.[24] Allerdings werden diese vermeintlich unentgeltlichen Leistungen in der Regel mit der Bereitstellung und Einwilligung in die Verarbeitung von personenbezogenen Daten „bezahlt". Es ist überzeugend, diese Entscheidung über das „Bezahlen mit Daten", das an die Stelle einer finanziellen Gegenleistung tritt, als geschäftliche Entscheidung i. S. v. § 2 Abs. 1 Nr. 9 UWG/Art. 2 lit. k) UGP-RL anzusehen, um der Gleichstellung dieser beiden Szenarien, die von § 312 Abs. 1a

---

[19] EuGH GRUR 2014, 196 Rn. 35 ff. – Trento Sviluppo.

[20] BGH GRUR 2016, 1073 Rn. 34 – Geo-Targeting; KG GRUR-RR 2019, 34 Rn. 26.

[21] BGH GRUR 2015, 698 Rn. 20 – Schlafzimmer komplett.

[22] So hat der BGH eine geschäftliche Entscheidung bejaht, wenn das Anlocken die Möglichkeit zur Mitgliederwerbung eröffnete, BGH GRUR 2008, 186 Rn. 28.

[23] BGH ZD 2014, 469 – Nordjob-Messe.

[24] Verneinend Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 2 Rn. 156a; a. A. *Omsels*, Die geschäftliche Entscheidung, WRP 2016, 553 (559).

BGB und § 327 Abs. 3 BGB auf Grundlage der Digitale-Inhalte-RL 2019/770 angestrebt wird, Rechnung zu tragen. Damit fallen also Dark Patterns, die den Verbraucher dazu veranlassen, im Gegenzug für einen unentgeltlich angebotenen Dienst Daten preiszugeben, in den Anwendungsbereich des durch die UGP-Richtlinie harmonisierten Lauterkeitsrechts.

2. Unzulässige Geschäftspraktiken

Um ein hohes Verbraucherschutzniveau sowie Rechtssicherheit[25] zu gewährleisten, erklärt die sog. „schwarze Liste" im Anhang I zur UGP-RL bzw. im Anhang zu § 3 Abs. 3 UWG zum einen bestimmte irreführende und zum anderen bestimmte aggressive Geschäftspraktiken für stets unzulässig. Diese Verbote gelten unabhängig von der Beurteilung des Einzelfalls anhand der in §§ 4a ff. UWG geregelten Unlauterkeitstatbestände[26] und somit unabhängig davon, ob eine Geschäftspraktik zu einer spürbaren Beeinträchtigung (so noch die Formulierung in § 3 Abs. 1 und 2 UWG a. F.) der Verbraucher- oder anderer geschützter Interessen führte[27] bzw. ob sie dazu geeignet ist, Verbraucher zu einer geschäftlichen Entscheidungen zu bewegen, die sie andernfalls nicht getroffen hätten.[28] Die Europäische Kommission ging davon aus, dass die in der Liste aufgeführten Praktiken „die Entscheidung eines Durchschnittsverbrauchers stets wesentlich beeinflussen und […] gegen das Gebot der beruflichen Sorgfalt" verstoßen.[29] Damit zählt die „schwarze Liste" Praktiken auf, deren Unlauterkeit so gravierend bzw. unabhängig von den konkreten Umständen gegeben ist, so dass sie per se untersagt werden. Die Verbote gelten also, ohne dass zu prüfen ist, ob die Geschäftspraktik im konkreten Einzelfall das wirtschaftliche Verhalten der Verbraucher beeinflusst.[30]

Nach der „schwarzen Liste" sind einige der eingangs genannten Beispiele für manipulative Designs als irreführende Geschäftspraktiken unzulässig. Dies gilt z. B. nach Nr. 7 für sog. Countdown Patterns.[31] Das Verbot erfasst allerdings nur unwahre Angaben über die (sehr knappe[32]) zeitliche Begrenzung des Angebots von Waren oder Dienstleistungen,[33] so dass eine vom Anbieter selbst tatsäch-

---

[25] S. Erw. 17 UGP-RL.

[26] Erw. 17 UGP-RL; EuGH GRUR 2022, 87 Rn. 67 – StWL/eprimo.

[27] Begr. RegE UWG 2008, BT-Drs. 16/10145, 30; *Alexander*, Die „Schwarze Liste" der UGP-Richtlinie und ihre Umsetzung in Deutschland und Österreich, GRUR Int 2010, 1025 (1030).

[28] Begr. RegE UWG 2008, BT-Drs. 16/10145, 30; Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 3 Rn. 4.7.

[29] Begründung zum Kommissionsentwurf (KOM (2003) 356 endg), Rn. 30 a. E.

[30] *Alexander*, GRUR Int 2010, 1025 (1027).

[31] *Martini/Kramme/Seeliger*, VuR 2022, 123 (125).

[32] Köhler/Bornkamm/Feddersen/*Köhler*, UWG, Anh. § 3 Rn. 7.6; *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 40.

[33] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (64); MüKoUWG/*Alexander*, UWG nach § 3 Abs. 3 Nr. 7 Rn. 16.

lich gesetzte kurze Angebotsfrist[34] oder der ebenfalls Knappheit suggerierende zutreffende Hinweis, dass sich auch andere Kunden gerade ein bestimmtes Angebot ansehen, damit nicht unter diese Regelung fallen.[35] Unzulässig sind auch das Bewerben von mit Kosten verbundener Waren oder Dienstleistungen als kostenlos (Nr. 21), sog. Hidden-Cost-Patterns,[36] sowie das Suggerieren einer nicht-bestehenden vertraglichen Zahlungspflicht durch die Übermittlung von Werbematerial bzw. die Lieferung unbestellter Waren (Nr. 22 und 29).[37] Im Rahmen der Änderungen zum 28.5.2022 kommen noch Verbote zu verdeckter Werbung in Suchergebnissen in Nr. 11a[38] sowie zu gefälschten oder ungeprüften Verbraucherbewertungen in Nr. 23b und 23c, die die sog. Social-Proof-Patterns[39] untersagen, hinzu. Durch die Reform in den Anhang zu § 3 Abs. 3 UWG als aggressive Geschäftspraktik verschoben wird auch das bislang in § 7 Abs. 2 Nr. 1 UWG geregelte Verbot des hartnäckigen und unerwünschten Ansprechens mittels für den Fernabsatz geeigneter Mittel der kommerziellen Kommunikation (Nr. 26). Als hartnäckig sah der EuGH bereits den Erhalt von drei Werbemails innerhalb eines Zeitraums von gut einem Monat an.[40] Die Regelung adressiert damit Nagging-Pattern, z. B. hartnäckige und wiederholte Aufforderungen zur Erteilung der Erlaubnis zur Datenerhebung.[41]

Zusammenfassend ist damit festzuhalten, dass bestimmte manipulative Designs durch den Anhang zu § 3 Abs. 3 UWG per se für unlauter erklärt werden. Auch wenn der Zusammenstellung der „schwarzen Liste" im Anhang der UGP-Richtlinie wohl keine systematische Analyse von Dark Patterns vorausging, ist es aber nicht allein dem Zufall geschuldet, dass sie nur die aufgeführten Dark Patterns untersagt.[42] Vielmehr ist diese begrenzte Auswahl auch darauf zurückzuführen, dass nicht alle manipulativen Designs per se untersagungswürdig und damit für eine Aufnahme in die „schwarze Liste" geeignet sind, da die Grenze zwischen zulässiger Beeinflussung und regulierungsbedürftiger Manipulation in vielen Konstellationen nur im jeweiligen Einzelfall zu ziehen ist.

Allerdings wäre es unter dem Gesichtspunkt der Rechtsicherheit sowie der Gewährleistung EU-weit harmonisierter Schutzstandards wünschenswert, wenn der europäische Gesetzgeber prüfen würde, ob es sinnvoll ist, den Anhang zur

---

[34] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (64); *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 40.

[35] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (64).

[36] *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 40.

[37] Kühling/Sauerborn, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 40.

[38] Diese Praktik wurde zuvor als irreführend iS von § 5 Ans. 2 Nr. 1 UWG a. F. eingeordnet, s. LG München I MMR 2016, 257 (259 f.).

[39] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (65).

[40] EuGH GRUR 2022, 87 Rn. 73 – StWL/eprimo.

[41] S. zu § 4a UWG *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (66); *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 43.

[42] So wohl *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 40; *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (65).

UGP-Richtlinie um weitere Dark Patterns zu erweitern.[43] In Betracht käme z. B., über den in Nr. 11 geregelten Fall der Nutzung von vermeintlich redaktionellen Inhalten hinaus die Tarnung von Werbung als Bestandteil der Benutzeroberfläche oder als Steuerungselement, damit Nutzer mit dem Inhalt interagieren (sog. Disguised Ads-Dark Pattern[44]), in die Liste aufzunehmen. Zwar besteht bereits nach § 6 Abs. 1 Nr. 1 TMG die Verpflichtung, kommerzielle Kommunikation klar als solche erkennbar zu machen. Diese Verpflichtung kann über den Rechtsbruchtatbestand des § 3a UWG auch lauterkeitsrechtlich durchgesetzt werden,[45] allerdings wiederum nur unter der Voraussetzung, dass der Verstoß zur spürbaren Interessenbeeinträchtigung geeignet ist.

3. Irreführungsverbote

Transparenzanforderungen ergeben sich des Weiteren aus dem Verbot irreführender geschäftlicher Handlungen gemäß §§ 5, 5a und § 5b UWG.

a) Irreführung

Irreführend sind sowohl unwahre Angaben als auch sonstige zur Täuschung geeignete Angaben (§ 5 Abs. 2 UWG). Damit erfasst die Regelung nicht nur das Bereitstellen falscher Informationen, sondern auch richtiger Informationen, die aber so aufbereitet sind, dass sie von der Zielgruppe falsch verstanden werden. Solche die Art und Weise der Informationsbereitstellung betreffende manipulative Designs können z. B. Countdown-Patterns sein, die zwar nicht fälschlicherweise die Knappheit der Waren behaupten, aber dies durch die Art und Weise der Darstellung suggerieren.[46]

Zwar besteht keine allgemeine Aufklärungspflicht des Unternehmers hinsichtlich möglicherweise relevanter Tatsachen.[47] Allerdings kann eine Irreführung eines Verbrauchers auch dadurch erfolgen, dass der Unternehmer wesentliche Informationen vorenthält, z. B. indem die Informationen verheimlicht oder in unklarer, unverständlicher oder zweideutiger Weise bereitgestellt werden (§ 5a Abs. 2 UWG). Unlauter kann hierbei auch eine die wesentlichen Informationen verschleiernde graphische Darstellung z. B. von Buttons oder Textelementen sein.[48]

Zu den bereitzustellenden Informationen gehört u. a. auch der kommerzielle Zweck einer geschäftlichen Handlung, sofern sich dieser nicht unmittelbar aus

---

[43] Ebenso *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (72).

[44] S. die Beschreibung auf der Seite des Dark Pattern Detection Project: https://dapde.de/de/dark-patterns/arten-und-beispiele/irref%C3%BChrung2/.

[45] BGH MMR 2021, 892 Rn. 89 – Influencer III.

[46] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (67); *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 44.

[47] BGH GRUR 2018, 541 Rn. 38 – Knochenzement II.

[48] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (67).

den Umständen ergibt (§ 5a Abs. 4 S. 1 UWG). Danach müsste also auch in dem eingangs erwähnten Beispiel des „Gesprächs" mit einem virtuellen Assistenten oder Chatbot, der einem Verbraucher zielgerichtet Produkteempfehlungen gibt, kenntlich gemacht werden, dass das vermeintlich fürsorgliche Gespräch kommerziellen Zwecken dient.[49]

b) Verbraucherleitbild

Von zentraler Bedeutung ist allerdings die Frage, nach welchem Maßstab zu beurteilen ist, ob eine Gestaltung irreführend und daher dazu geeignet ist, einen Verbraucher zu einer geschäftlichen Entscheidung zu veranlassen, die er andernfalls nicht getroffen hätte. Gemäß § 3 Abs. 4 UWG ist auf den durchschnittlichen Verbraucher abzustellen. Nur dann, wenn sich die geschäftliche Handlung an eine bestimmte Gruppe von Verbrauchern wendet, ist auf ein durchschnittliches Mitglied dieser Gruppe abzustellen und dabei ggf. auch eine besondere Schutzbedürftigkeit aufgrund von geistigen oder körperlichen Beeinträchtigungen, Alter oder Leichtgläubigkeit zu berücksichtigen. Damit ist der Durchschnittsadressat in jedem Einzelfall normativ zu bestimmen.

Ausgangspunkt ist die Erw. 18 S. 2 UGP-RL und der ständigen Rechtsprechung des EuGH[50] zugrundeliegende Vorstellung eines grundsätzlich „angemessen gut unterrichtet und angemessen aufmerksam und kritischen" durchschnittlichen Verbrauchers. Im Schrifttum wird in Zweifel gezogen, ob ein solches von genereller Rationalitätserwartung geprägtes Verbraucherleitbild eine angemessene Berücksichtigung der Eigenarten von manipulativen Designs zulässt.[51] Gerade bei unterschwellig wirkenden Designs, die gezielt kognitive Heuristiken ausnutzen, erscheint es durchaus wahrscheinlich, dass auch ein durchschnittlicher Verbraucher – trotz vollständiger Information – typischerweise nicht dazu in der Lage ist, rational seinen Entscheidungspräferenzen zu folgen.[52]

Allerdings gibt die UGP-RL neben dem reibungslosen Funktionieren des Binnenmarktes explizit auch das Ziel eines hohen Verbraucherschutzes vor (Art. 1 UGP-RL). Unter Bezugnahme hierauf[53] legt der EuGH in seiner Rechtsprechung in den letzten Jahren bei der Auslegung der im Anhang geregelten unzulässigen Geschäftspraktiken Maßstäbe an,[54] die auch die begrenzte Informationsverar-

---

[49] Ebenso bereits *Köbrich/Froitzheim*, Lass uns quatschen – Werbliche Kommunikation mit Chatbots, WRP 2017, 1188 Rn. 16 ff.

[50] EuGH GRUR Int. 1998, 795 Rn. 37 – Gut Springheide; EuGH GRUR 2011, 930 Rn. 23 – Ving Sverige; EuGH GRUR 2022, 577 Rn. 47 – Upfield Hungary; EuGH GRUR 1993, 747 Rn. 17 – Yves Rocher.

[51] *Weinzierl*, NVwZ 2020-Extra, 1 (8).

[52] So ganz zu Recht MüKoUWG/*Raue*, UWG § 4a Rn. 17.

[53] S. auch EuGH GRUR 2012, 1269 Rn. 48 – Purely Creative.

[54] EuGH GRUR 2012, 1269 – Purely Creative: Unlauter gemäß Nr. 31 ist das Ausnutzen der durch die Mitteilung des Gewinns eines Preises ausgelösten psychologischen Wirkung (Rn. 38), die

beitungskapazitäten und Informationsdefizite von Verbrauchern in der konkreten Situation berücksichtigen und damit von dem Leitbild des gut informierten und aufmerksamen Durchschnittsverbrauchers abweichen.[55] Ob die Gefahr einer Irreführung vorliegt, ist daher durch eine Interessenabwägung zwischen den Freiheitsrechten des Wettbewerbers und dem allgemeinen Interesse an freiem Wettbewerb, andererseits dem Interesse der Allgemeinheit an einem unverfälschten Wettbewerb unter Berücksichtigung der Umstände des konkreten Einzelfalls festzustellen.[56] Das normative Verbraucherleitbild ist damit hinreichend flexibel und entwicklungsoffen, um neben dem Kriterium der situationsadäquat gebotenen Aufmerksamkeit[57]auch die Auswirkungen von unterschwellig wirkenden manipulativen Designs, die gezielt kognitive Heuristiken ausnutzen, im Rahmen der Gesamtbetrachtung einzubeziehen.[58]

4. Weitere Unlauterkeitstatbestände

Neben dem Verbot bestimmter Geschäftspraktiken durch § 3 Abs. 3 UWG i. V. m. dem Anhang sowie den aus den Irreführungsverboten resultierenden Transparenzvorgaben kommen zudem die Verbotstatbestände der § 4a UWG und § 7 UWG sowie der Rechtsbruchtatbestand des § 3a UWG und die Generalklausel des § 3 Abs. 2 in Betracht.

a) Verbot aggressiver geschäftlicher Handlungen, § 4a UWG

§ 4a UWG untersagt aggressive geschäftliche Handlungen, die geeignet sind, den Verbraucher zu einer geschäftlichen Entscheidung zu veranlassen, die dieser andernfalls nicht getroffen hätte. Die Regelung definiert eine geschäftliche Handlung als aggressiv, wenn sie dazu geeignet ist, die Entscheidungsfreiheit des

---

den Verbraucher zu einer nicht rationalen Entscheidung veranlassen kann, indem er den schnellsten Weg (z. B. Anruf einer Mehrwertnummer oder aufwändige Fahrt) wählt, um in Erfahrung zu bringen, welchen Preis er gewonnen hat, obwohl gerade dieser vielleicht zu den höchsten Kosten führt (Rn. 50). Damit fließt in die Bewertung ein, welcher Grad an Aufmerksamkeit angesichts der Art und Weise der Ansprache situationsadäquat ist; ähnlich schon BGH GRUR 2000, 619 (621) – Orient-Teppichmuster. Vgl. nun auch die Entscheidung des EuGH GRUR 2022, 87 Rn. 43 – StWL/eprimo, in der der Gerichtshof bei einer grau unterlegten, mit dem Hinweis „Anzeige" versehenen Werbeeinblendung im E-Mail-Postfach trotz dieser Ausgestaltung die „Gefahr einer Verwechslung" mit einer E-Mail sah und ein Einwilligungserfordernis nach Art. 13 E-Privacy-RL 2002/58 bejahte (der allerdings nicht wirtschaftliche Verbraucherinteressen, sondern gem. Art. 1 E-Privacy-RL das Recht auf Privatsphäre und Datenschutz schützt).

[55] Harte-Bavendamm/Henning-Bodewig/*Glöckner*, UWG, 5. Aufl. 2021, Einleitung: B: Europäisches Lauterkeitsrecht Rn. 442.

[56] Harte-Bavendamm/Henning-Bodewig/*Glöckner*, UWG, Einleitung: B: Europäisches Lauterkeitsrecht Rn. 466 ff.

[57] Ohly/Sosnitza/*Sosnitza*, Gesetz gegen den unlauteren Wettbewerb, 7. Aufl. 2016, UWG § 2 Rn. 123.

[58] So ganz zu Recht MüKoUWG/*Raue*, UWG, 3§ 4a Rn. 17.

Verbrauchers durch Belästigung, Nötigung oder unzulässige Beeinflussung erheblich zu beeinträchtigen, § 4 Abs. 1 S. 2 UWG. Schutzzweck der Regelung ist die Sicherstellung der Entscheidungsfreiheit des Verbrauchers (Erw. 16 UGP-RL). § 4a Abs. 2 UWG zählt nicht abschließend[59] Kriterien auf, die für die Feststellung, ob eine geschäftliche Handlung aggressiv ist, zu berücksichtigen sind. Genannt wird unter anderem die Behinderung eines Verbrauchers an der Ausübung vertraglicher Rechte, worunter z. B. sog. roach motel pattern fallen können.[60]

Unter dem autonom und einheitlich auszulegenden,[61] aber in der Richtlinie nicht definierten Begriff der Belästigung gemäß § 4a Abs. 1 S. 2 Nr. 1 UWG wird im Schrifttum ein störender Eingriff in die Privatsphäre verstanden, der die „Grenzen des sozialadäquaten Umgangs" überschreitet und dem sich der Verbraucher nicht ohne weiteres, zum Beispiel durch Wegsehen, Weghören oder Weggehen, entziehen kann.[62] Als Beispiele genannt werden unerbetene Hausbesuche sowie das unerbetene und hartnäckige Zusenden von Nachrichten mittels Telefon, Fax oder E-Mail, wobei aber jeweils zu prüfen ist, ob die jeweiligen Umstände, insbesondere die in § 4a Abs. 2 UWG genannten, dazu führen, dass hiermit auch eine wesentliche Beeinflussung der Entscheidungsfreiheit einhergeht.[63] Das unerbetene Aufschalten eines zusätzlichen Dienstes zu den einem Bestandskunden bislang erbrachten Leistungen ohne zusätzliche Kosten und bei Gewährung einer Opt-out-Möglichkeit durch ein „einfach auszuübendes Widerrufsrecht" stellt nach Ansicht des BGH hingegen aufgrund der einfachen Ausweichmöglichkeit keine Belästigung dar.[64] In seiner früheren Rspr. sah der BGH z. B. das unerbetene Ansprechen durch einen als solchen erkennbaren Werber in der Öffentlichkeit in der Regel nicht als Belästigung an, wenn sich der Angesprochene der Ansprache ohne weiteres durch Weitergehen entziehen konnte.[65]

Unter den Tatbestand der Belästigung können z. B. sog. Nagging-Patterns, z. B. hartnäckige und wiederholte Aufforderung zur Erteilung der Erlaubnis zur Datenerhebung fallen.[66] Vom jeweiligen Einzelfall abhängig ist hingegen die Beurteilung von Pop-Up-Fenstern auf einer kostenlos zugänglichen Internetseite, die Teile des Inhalts überdecken und Besucher dazu nötigen, die Pop-Up-Fenster wegzuklicken. Grundsätzlich ist diese Situation nicht mit der Zusendung unerwünschter Werbemails vergleichbar, da das Anzeigen von Pop-Ups auf frei zugänglichen

---

[59] Ohly/Sosnitza/*Sosnitza*, UWG, § 4a Rn. 147.

[60] *Kühling*/*Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 42.

[61] Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 4a Rn. 1.37.

[62] Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 4a Rn. 1.40 ff.; MüKoUWG/*Raue*, UWG § 4a Rn. 114 ff.

[63] Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 4a Rn. 1.42 ff.

[64] BGH GRUR 2019, 750 Rn. 34 – WifiSpot.

[65] BGH GRUR 2005, 443 (444 f.) – Ansprechen in der Öffentlichkeit II; zustimmend auch für § 4a UWG Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 4a Rn. 1.43.

[66] *Martini*/*Drews*/*Seeliger*/*Weinzierl*, ZfDR 2021, 47 (66); *Kühling*/*Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 43.

Internetseiten nicht in gleicher Weise in die Privatsphäre des Nutzers eingreift, dem es frei steht, den Besuch der Internetseite zu beenden.[67] Zudem erwartet der Nutzer, anders als im Falle eines passwortgeschützten E-Mail-Postfachs,[68] bei einer allgemein zugänglichen Internetseite nicht nur individuell an ihn gerichtete Nachrichten. Anders zu beurteilen wären hingegen Konstellationen, in denen die Pop-Ups nicht einfach weggeklickt werden können[69] oder wenn der Verbraucher auf den Besuch der Seite angewiesen ist, um z. B. einen Social-Media Account zu bedienen.

Eine unzulässige Beeinflussung gemäß § 4a Abs. 1 S. 2 Nr. 3 UWG liegt dann vor, wenn der Unternehmer eine Machtposition gegenüber dem Verbraucher im Rahmen einer unzulässigen Beeinflussung zur Ausübung von Druck ausnutzt, so dass die Fähigkeit des Verbrauchers zu einer informierten Entscheidung wesentlich eingeschränkt wird (§ 4a Abs. 1 Satz 3 UWG/Art. 2 lit. j) UGP-RL. Dieses Verbot könnte z. B. für Fälle des Confirmshaming relevant werden, in denen Unternehmer die Angst von Verbrauchern vor Isolation oder sozialer Ächtung ausnutzen, um diese zum Erwerb eines Produkts zu motivieren.[70] Allerdings stellt sich die Frage, wann der mittels manipulativen Designs erzeugte Druck so intensiv ist, dass eine unzulässige Beeinflussung vorliegt. Nach bisheriger Auslegung würde dies voraussetzen, dass der Handelnde den Eindruck erweckt, der Verbraucher müsse mit irgendwelchen Nachteilen außerhalb des angestrebten Geschäfts rechnen, falls er die von ihm erwartete geschäftliche Entscheidung nicht trifft.[71] Auch die in § 4a Abs. 2 UWG genannten Umstände belegen, dass der Gesetzgeber durchaus an Situationen gedacht hat, in denen ein erhebliches Maß an Druck ausgeübt wird. Entsprechend wird auch im Schrifttum zutreffend darauf hingewiesen, dass die Regelung Verbraucher nicht vor jeder Form von sozialem Druck bewahren soll.[72]

Denkbar ist, dass auch die Ansprache durch einen vermeintlich empathischen Chatbot zu einem vergleichbaren sozialen Druck führt. Zwar lässt sich hier einwenden, dass sich der Adressat der Ansprache durch einen Klick leicht entziehen könne und dass die Beziehung des Adressaten zu dem Chatbot eine andere Qualität habe als zu einem engagierten Verkäufer, so dass hier kein vergleichbarer physischer oder moralischer Kaufzwang entstehe. Andererseits gibt es zumindest anekdotische Evidenz dafür, dass es Situationen gibt, in denen Menschen wider

---

[67] A. A. wohl *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (66); vgl. zu § 7 UWG auch *Mankowski* in Fezer/Büscher/Obergfell, Lauterkeitsrecht: UWG, 3. Aufl. 2016, Wettbewerbsrecht des Internets (S 12), Rn. 149 ff. m. w. N.

[68] EuGH GRUR 2022, 87 Rn. 69 – StWL/eprimo.

[69] S. zu § 7 UWG OLG Köln MMR 2014, 51; KG MMR 2014, 44; LG Düsseldorf MMR 2003, 486; LG Berlin GRUR-RR 2011, 332.

[70] *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 48.

[71] OLG München GRUR 2017, 1147 Rn. 190; Köhler/Bornkamm/Feddersen/*Köhler*, UWG § 4a Rn. 1.59.

[72] MüKoUWG/*Raue*, UWG § 4a Rn. 12 f.

besseres Wissen auch virtuellen Assistenten und Robotern die Eigenschaften von Lebewesen zuschreiben.[73] Insofern lässt sich nicht grundsätzlich ausschließen, dass zu einer Situation kommen kann, in der diese Art von Beeinflussung die Entscheidungsfreiheit eines Verbrauchers erheblich beeinträchtigen kann.[74]

Große Bedeutung kommt somit auch hier der Frage zu, wann ein Dark Pattern, wie von § 4a Abs. 1 S. 2 UWG vorausgesetzt, im konkreten Fall unter Berücksichtigung aller Umstände dazu geeignet ist, die Entscheidungsfreiheit des Verbrauchers erheblich zu beeinträchtigen. Wie bereits dargestellt, ist hierbei auf den durchschnittlich aufmerksamen, informierten und angemessen kritischen Adressaten abzustellen (§ 3 Abs. 4 UWG). Bei der Beurteilung, ob die Entscheidungsfreiheit durch eine bestimmte aggressive geschäftliche Handlung erheblich beeinträchtigt ist, wird man aber nicht allein auf die Fähigkeiten zum Erfassen der Entscheidungsgrundlage abstellen müssen, sondern auch auf die Art und Weise des Entscheidungsprozesses. Entscheidende Kriterien sind daher auch die Beeinflussbarkeit, Empfindlichkeit, Widerstandsfähigkeit und Willenskraft eines durchschnittlich beeinflussbaren Adressaten.[75] Gerade bei unterschwellig wirkenden manipulativen Designs, die gezielt kognitive Heuristiken ausnutzen, erscheint es durchaus wahrscheinlich, dass auch ein durchschnittlicher Verbraucher – trotz vollständiger Information – typischerweise nicht dazu in der Lage ist, rational seinen Entscheidungspräferenzen zu folgen, und daher in seiner Entscheidungsfreiheit erheblich beeinträchtigt wird.[76]

b) Verbot unzumutbarer Belästigungen, § 7 UWG

Ergänzend zu § 4a Abs. 1 S. 2 Nr. 1 UWG untersagt § 7 UWG unzumutbare Belästigungen, insbesondere durch Werbung. Die für Nagging-Patterns relevante frühere Regelung des § 7 Abs. 2 Nr. 1 wurde nun in den Anhang zu § 3 Abs. 3 UWG als aggressive Geschäftspraktik verschoben (Nr. 26). Geklärt ist mittlerweile auch, dass das Einblenden von Werbenachrichten im E-Mail-Postfach ohne Einwilligung gegen den durch § 7 Abs. 2 Nr. 3 UWG umgesetzten Art. 13 E-Privacy-RL 2002/58 verstößt, wenn die Gefahr einer Verwechslung mit E-Mails besteht.[77] Auch die Kontaktaufnahme durch einen z. B. zu Werbeansprachen eingesetzten Chatbot ist nach dieser Regelung wohl nur mit Zustimmung des Adressaten zulässig.[78]

---

[73] *Darling*, ‚Who's Johnny?' – Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy, in: Lin/Bekey/Abney/Jenkins, ROBOT ETHICS 2.0, 2017; abrufbar unter https://ssrn.com/abstract=2588669.

[74] *Köbrich/Froitzheim*, WRP 2017, 1188 Rn. 13 ff.

[75] MüKoUWG/*Raue*, UWG § 4a Rn. 70.

[76] So ganz zu Recht MüKoUWG/*Raue*, 3UWG § 4a Rn. 17.

[77] EuGH GRUR 2022, 87 Rn. 32 ff. – StWL/eprimo.

[78] *Köbrich/Froitzheim*, WRP 2017, 1188 Rn. 10.

c) Rechtsbruchtatbestand und Verbrauchergeneralklausel

Nur kurz hinzuweisen ist noch auf den Rechtsbruchtatbestand des § 3a UWG sowie die Generalklausel des § 3 Abs. 2 UWG als Auffangtatbestand. Mit dem Rechtsbruchtatbestand können Verstöße gegen Regelungen außerhalb des Lauterkeitsrechts erfasst werden, die iSd § 3a UWG auch das Marktverhalten im Interesse der Marktteilnehmer regeln. Allerdings muss auch nach diesem Unlauterkeitstatbestand der Verstoß geeignet sein, die Interessen von Verbrauchern, sonstigen Marktteilnehmern oder Mitbewerbern spürbar zu beeinträchtigen. Dadurch soll die Verfolgung von Bagatellfällen ausgeschlossen werden, weshalb die Schwelle nach den Vorstellungen des Gesetzgebers nicht zu hoch anzusetzen sein soll.[79] Anders als bei §§ 4a, 5 und 5a UWG kommt es für die Spürbarkeit i. S. d. § 3a UWG aber nicht darauf an, ob der Verstoß auch geeignet ist, den Verbraucher zu einer geschäftlichen Entscheidung zu veranlassen, die er sonst nicht getroffen hätte. Vielmehr ist allein entscheidend, ob der Verstoß geeignet ist, durch die gesetzliche Regelung geschützte Interesse, wie z. B. im Fall von § 6 TMG an der Erkennbarkeit der kommerziellen Kommunikation, zu beeinträchtigen.

Schließlich kommt für manipulative Designs auch der Auffangtatbestand der „Verbrauchergeneralklausel" in § 3 Abs. 2 UWG in Betracht. Danach sind geschäftliche Handlungen, die sich an Verbraucher richten oder diese erreichen, unlauter, wenn sie nicht der unternehmerischen Sorgfalt entsprechen und dazu geeignet sind, das wirtschaftliche Verhalten des Verbrauchers wesentlich zu beeinflussen. Auch hier ist wieder von maßgeblicher Bedeutung, ob manipulative Ausgestaltungen in Bezug auf den durchschnittlichen Verbraucher hinreichend effektiv sind.

5. Zwischenergebnis

Zusammenfassend ist festzuhalten, dass sich das Lauterkeitsrecht im Hinblick auf Dark Patterns als hinreichend entwicklungsoffen gezeigt hat, um nicht nur wirtschaftliche Interessen der Verbraucher, sondern auch andere Interessen, z. B. Privatheit und Autonomie, zu schützen. Dies gilt nicht nur für das deutsche UWG, sondern durch die weite Auslegung ihres Anwendungsbereichs durch den EuGH auch für die UGP-Richtlinie.

Angesichts der Herausforderung, im Spannungsfeld von Verbraucherschutz und unternehmerischer Freiheit regelungsbedürftige Manipulation von der noch sozialadäquaten Beeinflussung abzugrenzen, hat sich die Regelungstechnik bewährt, Geschäftspraktiken, die unabhängig vom Einzelfall als unlauter zu beurteilen sind, durch spezifische Regelungen zu untersagen und für andere Geschäftspraktiken, deren Zulässigkeit nur im Einzelfall zu bewerten ist, offene, wertungsabhängige Tatbestände zu schaffen, die dann von den Gerichten auszuformen sind. Während

[79] Begr. RegE UWG 2004, BT-Dr 15/1487, 17; BGH GRUR 2008, 186 Rn. 25 – Telefonaktion.

für irreführende Dark Patterns mit den Regelungen §§ 5 ff. UWG eine hinreichend ausdifferenzierte Regelung vorliegt, bleiben bei der Subsumtion von Nagging Patterns und Patterns, die ein emotionales Framing vornehmen, z. B. Confirmshaming Patterns, unter den Tatbestand der aggressiven Geschäftspraktiken i. S. v. § 4a UWG allerdings Unklarheiten, wann eine hinreichend schwerwiegende Belästigung oder unzulässige Beeinflussung vorliegt. Von entscheidender Bedeutung im Rahmen der genannten Tatbestände sowie auch der Verbrauchergeneralklausel des § 3 Abs. 2 UWG ist das zugrundgelegte Verbraucherleitbild. Der von der Rechtsprechung angelegte normative Maßstab des Durchschnittsverbrauchers ist zwar hinreichend flexibel und entwicklungsoffen, um neben dem Kriterium der situationsadäquat gebotenen Aufmerksamkeit auch die Auswirkungen von unterschwellig wirkenden manipulativen Designs, die gezielt kognitive Heuristiken ausnutzen, im Rahmen der Gesamtbetrachtung einzubeziehen. Eine weitere Herausforderung ergibt sich allerdings daraus, dass den mit der Rechtsdurchsetzung betrauten Institutionen – im Gegensatz zu den Verwendern – häufig die empirischen Daten fehlen, um zu bestimmen, welchen Dark Patterns in welchem Kontext ein hoher Grad an Manipulationseignung zukommt. Erwägenswert ist daher, Offenlegungsansprüche in Bezug auf entsprechende Studienergebnisse gegenüber den Verwendern zu schaffen.[80] Zudem ist ein Transfer verhaltenswissenschaftlicher Befunde in die Rechtsprechung und Rechtswissenschaft erforderlich.

III. Datenschutzrecht

Weitere Transparenzvorgaben ergeben sich aus dem Datenschutzrecht. Allerdings ist der Anwendungsbereich des Datenschutzrechts nur dann eröffnet, wenn eine Verarbeitung personenbezogener Daten stattfindet – wenn also Dark Patterns dazu eingesetzt werden, um Verbraucher zu einer Einwilligung in die Verarbeitung ihrer personenbezogenen Daten zu veranlassen oder wenn Daten verarbeitet werden, um Dark Patterns zielgerichtet an konkrete Adressaten anzupassen. Hingegen ist der Anwendungsbereich des Datenschutzrechts nicht eröffnet, wenn manipulative Designs dazu eingesetzt werden, um Verbraucher z. B. zum Abschluss eines Kauf- oder Lizenzvertrags zu motivieren, der keine über Art. 6 Abs. 1 lit. b) DSGVO hinausgehende Datenverarbeitung beinhaltet. Insofern ist der Anwendungsbereich des Datenschutzrechts im Kontext von manipulativen Designs wesentlich begrenzter als der des Lauterkeitsrechts.[81]

Um die informationelle Autonomie des Einzelnen zu gewährleisten, ist Transparenz ein zentraler Grundsatz der DSGVO (Art. 5 Abs. 1 lit. a) DSGVO), der u. a. durch die Informationspflichten der Art. 12 ff. DSGVO sowie das Erfordernis der

[80] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (73).
[81] So bereits *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (59).

Informiertheit als Wirksamkeitsvoraussetzung für die Einwilligung konkretisiert wird.

1. Einwilligung

Die Rechtmäßigkeit der Verarbeitung personenbezogener Daten setzt gemäß Art. 6 Abs. 1 DSGVO voraus, dass sie auf einer Einwilligung oder einem anderen der genannten gesetzlichen Erlaubnistatbestände beruht. Hierbei kommt der bereits primärrechtlich durch Art. 8 Abs. 2 GRCh vorgegebenen Einwilligung als Ausdruck des Selbstbestimmungsrechts der betroffenen Person eine besondere Bedeutung zu,[82] wenngleich sie in der Systematik des Art. 6 Abs. 1 DSGVO gleichrangig neben den anderen gesetzlichen Erlaubnistatbeständen steht.[83]

Allerdings stellt sich angesichts der für die Teilnahme an digitaler Kommunikation erforderlichen Anzahl von Einwilligungserklärungen die Frage, ob die Einwilligung als Rechtsinstrument dieser Funktion, das Selbstbestimmungsrecht des Einzelnen zu wahren, noch in vollem Maße gerecht wird.[84] Zu Recht wird darauf hingewiesen, dass der Gesetzgeber über eine „Einschätzungs- und Wertungsprärogative hinsichtlich der Frage (verfüge), welchen Anforderungen eine kommunikative Handlung zu genügen (habe), um als selbstbestimmt zu gelten und damit als Einwilligung im Rechtssinne eine Verarbeitung legitimieren zu können".[85] Insbesondere im Kontext manipulativer Designs ist daher zu untersuchen, unter welchen Voraussetzungen das für die Wirksamkeit der Einwilligung erforderliche Maß an Autonomie nicht mehr gegeben ist. Anders als nun der California Privacy Rights Act (CPRA) von 2020, der bestimmt, dass eine durch Dark Patterns erreichte Zustimmung keine wirksame Einwilligung ist,[86] adressiert die DSGVO Dark Patterns nicht explizit. Daher stellt sich die Frage, welche Auswirkungen der Einsatz manipulativer Designs auf die Wirksamkeitsvoraussetzungen, d. h. die Informiertheit, die Freiwilligkeit und das Vorliegen einer eindeutigen und bestätigenden Handlung hat.

a) Informiertheit

Eine wirksame Einwilligung erfordert gemäß Art. 4 Nr. 11 DSGVO eine informierte Entscheidung, damit die betroffene Person die Tragweite dieser Entscheidung erkennen kann. Um eine Informiertheit der betroffenen Person zu gewährleisten, muss die betroffene Person zumindest über die Identität des Verantwortlichen, die

---

[82] Simitis/Hornung/Spiecker/*Schantz*, Datenschutzrecht, DS-GVO Art. 6 Abs. 1 Rn. 11.

[83] Kühling/Buchner/*Buchner/Kühling*, DS-GVO BDSG, 3. Aufl. 2020, DSGVO Art. 7 Rn. 16; Ehmann/Selmayr/*Heckmann/Paschke*, DS-GVO, 2. Aufl. 2018, Art. 7 Rn. 19.

[84] Kühling/Buchner/*Buchner/Kühling*, DS-GVO Art. 7 Rn. 4, 10.

[85] Simitis/Hornung/Spiecker/*Klement*, Datenschutzrecht, DS-GVO Art. 7 Rn. 3.

[86] Cal. Civ. Code § 1798.140(h).

Verarbeitungszwecke, die Art der verarbeiteten Daten, das Widerrufsrecht und ggf. eine automatisierte Entscheidungsfindung i. S. v. Art. 22 DSGVO und Drittstaatentransfers informiert werden.[87] Zudem müssen die Informationen „in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache" zur Verfügung gestellt werden (vgl. Art. 12 DSGVO und Art. 7 Abs. 2 DSGVO). Hieran fehlt es z. B., wenn relevante Informationen durch ihre grafische Gestaltung nicht hinreichend wahrnehmbar (z. B. wegen zu kleiner oder nicht kontrastreicher Schrift) oder missverständlich formuliert sind.[88] Dies ist u. a. dann der Fall, wenn die Aufbereitung dazu geeignet ist, den Nutzer über die zur Verfügung stehenden Optionen zu täuschen, z. B. wenn im Rahmen der nach § 25 Abs. 1 S. 1 TTDSG erforderlichen Erteilung der Einwilligung zur Speicherung von Cookies auf Endgeräten die Auswahloption, die Einwilligung zu modifizieren oder zu verweigern, so dass nur notwendige Cookies verwendet werden dürfen, nicht als anklickbare Schaltfläche und damit nicht als gleichwertige Auswahloption zu erkennen ist.[89]

Weniger klar ist hingegen die Beurteilung von Gestaltungen, bei denen die (vollständigen und richtigen) Informationen in einer Weise dargestellt werden, die die Entscheidungsfindung zugunsten der Erteilung der Einwilligung beeinflusst, ohne jedoch unwahre oder täuschende Angaben zu enthalten. Hier kommt es auf den jeweiligen Einzelfall an, ob eine solche Darstellungsweise im Einzelfall aufgrund ihrer – unter Marketing-Gesichtspunkten durchaus angestrebten – Suggestivkraft zur Unwirksamkeit der Einwilligung führt. Eingehend diskutiert wird dies für die farbige oder größenmäßige Hervorhebung des Einwilligungs-Buttons im Rahmen von Cookie-Bannern, so dass die Einwilligungsoption der betroffenen Person als erste „ins Auge springt" und hierdurch die Wahrscheinlichkeit erhöht wird, dass diese ohne weitere Prüfung ihre Einwilligung erteilt.[90]

---

[87] *EDSA*, Leitlinien 05/2020 zur Einwilligung gemäß Verordnung 2016/679, S. 17 f. Es spricht vieles dafür, dass die umfangreichen Kataloge der bereitzustellenden Informationen nach Art. 13 und 14 DSGVO, die den Grundsatz der Transparenz und Fairness nach Art. 5 Abs. 1a) DSGVO umsetzen (so EuGH NJW 2019, 3433 Rn. 79 – Planet 49), zwar Anhaltspunkte liefern, jedoch nicht vollständig auf den Kontext der Informiertheit der Einwilligung übertragbar sind (Kühling/Buchner/*Buchner/Kühling*, DS-GVO Art. 7 Rn. 59; Gola DS-GVO/*Schulz*, DS-GVO, 2. Aufl. 2018, DS-GVO Art. 7 Rn. 36; a. A. Wolff/Brink (Hrsg.), BeckOK Datenschutzrecht/*Stemmer*, 39. Ed. 1.11.2021, DS-GVO Art. 7 Rn. 58).

[88] *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (56).

[89] LG Rostock ZD 2021, 166 Rn. 54; BeckOK DatenschutzR/*Stemmer*, DS-GVO Art. 7 Rn. 64; *M. Becker* CR 2021, 230 (237 f.); *Lehr/Dietmann/Krisam/Volkamer*, Manipulative Designs von Cookies, DuD 2022, 296 (298). Zwar gelten die Regelungen der DSGVO wegen der nach Art. 95 DSGVO bestehenden Sperrwirkung von Art. 5 Abs. 3 ePrivacy-RL als nicht direkt anwendbar. Da der Verweis durch diese Vorschrift auf die Datenschutz-RL 95/46 aber gem. Art. 94 Abs. 2 DS-GVO als Verweisung auf die DSGVO zu verstehen ist, sind die Normen der DSGVO insoweit heranzuziehen, LG Rostock ZD 2021, 166 Rn. 38.

[90] Für Unwirksamkeit bei Hervorhebung nur der Einwilligungsoption u. a. die französische Datenschutzbehörde *CNIL*, Délibération n° 2020-091 v. 17.9.2020, Rn. 17 f., abrufbar unter: https://

Hier stellt sich die bereits oben im lauterkeitsrechtlichen Kontext angesprochene Frage, welches Maß an Aufmerksamkeit von einem Nutzer gegenüber manipulativen Designs zu erwarten ist bzw. gesetzlich gefordert wird. Anders als Erw. 18 UGP-RL lassen sich der DSGVO keine expliziten Vorgaben entnehmen. Der im Singular formulierte Wortlaut von Art. 12 DSGVO spricht für eine an den jeweiligen Adressaten angepasste Informationsaufbereitung, was allerdings bei an eine Vielzahl von betroffenen Personen gerichteten Informationen in der Praxis nicht umsetzbar ist. Einen strengen Maßstab hat der EuGH in der sowohl zur DS-RL 95/46 als auch zu Art. 4 Nr. 11 DSGVO ergangenen Orange Romania Entscheidung angelegt,[91] was aber auch den konkreten Umständen des Fall geschuldet gewesen sein mag. Im Schrifttum wird zum Teil auf die Aufnahme- und Verarbeitungsfähigkeiten eines Durchschnittsadressaten abgestellt,[92] z. T. eine an das niedrigste Verständnisniveau angepasste Darstellung gefordert.[93] Angesichts der Bedeutung der Informiertheit als Wirksamkeitsvoraussetzung der Einwilligung zur Legitimation eines Eingriffs in das Recht auf Datenschutz erscheint die letztgenannte Ansicht überzeugender. Auch wenn die Beeinträchtigung der Möglichkeit zur Information bzw. zu einer freien Entscheidungsfindung durch die betroffenen Personen allein aufgrund der unterschiedlichen farbigen Gestaltung der Buttons eines Cookie-Banners nur relativ geringfügig beeinträchtigt wird, spricht somit angesichts des für den Einzelnen unverhältnismäßigen Aufwands, der mit einer Analyse jeder einzelnen Cookie-Abfrage einhergeht, und der asymmetrisch verteilten Macht zur Ausgestaltung der Benutzeroberflächen vieles dafür, aufgrund einer solchen Ausgestaltung der Cookie-Banner erteilte Einwilligungen als unwirksam einzuordnen.

Unklar ist auch, ob ein Framing der Informationen, z. B. ein Verweis auf die Knappheit eines Angebots oder ein Entscheidungscountdown, gegen die Ver-

---

www.cnil.fr/sites/default/files/atoms/files/lignes_directrices_de_la_cnil_sur_les_cookies_et_autres_traceurs.pdf; die irische Datenschutzbehörde *DPC*, Guidance Note: Cookies and other tracking technologies, S. 9, abrufbar unter: https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Guidance%20note%20on%20cookies%20and%20other%20tracking%20technologies.pdf; BeckOK DatenschutzR/*Stemmer*, DS-GVO Art. 7 Rn. 64.1.; *Rauer*/*Ettig*, Update Cookies 2020, ZD 2021, 18 (22); *Sesing*, Cookie-Banner – Hilfe, das Internet ist kaputt!, MMR 2021, 544 (547), der auch für klare Gestaltungsvorgaben zur Vereinheitlichung und zum Schutz der Nutzer vor einem information overload de lege ferenda plädiert; in der Tendenz auch LG Rostock ZD 2021, 166 Rn. 54, das allerdings primär darauf abstellt, dass die Option der Ablehnung/Modifikation im zu entscheidenden Fall nicht als anklickbare Schaltfläche zu erkennen war; für grds. Unbeachtlichkeit der Hervorhebung *Baumgartner*/*Hansch*, Onlinewerbung und Real-Time-Bidding, ZD 2020, 435 (437); *Ettig*/*Herbrich*, K&R 2020, 719 (721); *Kühling*/*Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 53.

[91] EuGH ZD 2021, 89 Rn. 40 – Orange Romania/ANSPDCP.

[92] Kühling/Buchner/*Buchner*/*Kühling*, DS-GVO, Art. 7 Rn. 60; Kühling/Buchner/*Bäcker*, DS-GVO Art. 12 Rn. 11; Gola/*Franck*, DS-GVO Art. 12 Rn. 22; wohl auch *EDSA*, Leitlinien 05/2020 zur Einwilligung gemäß Verordnung 2016/6790 5/2020, S. 18 f.

[93] Ehmann/Selmayr/Heckmann/Paschke, DS-GVO Art. 12 Rn. 17 f.; Taeger/Gabel/*Taeger*, DSGVO, 4. Aufl. 2022, Art. 7 Rn. 68; Taeger/Gabel/*Arning*/*Rothkegel*, DSGVO Art. 4 Rn. 344.

pflichtung aus Art. 12 DSGVO verstoßen kann. Im Schrifttum wird dies mit der Begründung abgelehnt, dass Art. 12–14 DSGVO durch die Verpflichtung, verständliche Informationen zur Verfügung zu stellen, darauf abzielten, einem Übermaß an Informationen sowie unverständlichen Informationen vorzubeugen. Im Falle eines solchen Framing seien die Informationen hingegen verständlich, der Mensch verarbeitet sie nur nicht rational.[94] Auch wenn die Informiertheit der Entscheidung als Wirksamkeitsvoraussetzung durch eine solche Gestaltung nicht in Frage gestellt sein mag, spricht jedoch vieles dafür, dass sie grundsätzlich die Freiwilligkeit der Entscheidung beeinträchtigen kann.

b) Freiwilligkeit

Eine Einwilligung ist nach Erw. 42 a. E. nur dann freiwillig, wenn die betroffene Person „eine echte oder freie Wahl hat und somit in der Lage ist, die Einwilligung zu verweigern oder zurückzuziehen, ohne Nachteile zu erleiden". Zudem zeigt Erw. 43, dass bei der Beurteilung der Freiwilligkeit auch zu berücksichtigen ist, ob ein klares Ungleichgewicht zwischen Verantwortlichem und betroffener Person besteht. Dies gilt nicht nur für das Verhältnis zwischen Arbeitgeber und Arbeitnehmer oder der öffentlichen Hand und den Bürgern, sondern kann im jeweils zu prüfenden Einzelfall auch das Verhältnis von Unternehmern und Verbrauchern einbeziehen.[95] Dies spricht dafür, dass hier auch das Machtungleichgewicht zu berücksichtigen ist, das sich aus der einseitigen Gestaltungsmacht des Anbieters über die Ausgestaltung der Kommunikation und damit zugleich der Entscheidungsarchitektur ergeben kann.

Zwar besteht Einigkeit darüber, dass nicht jeder noch so geringe Nachteil, der mit der Verweigerung der Einwilligung verbunden ist, schon die Freiwilligkeit entfallen lässt.[96] Ungeklärt ist allerdings, wie schwerwiegend ein Nachteil sein muss, damit dies der Fall ist.[97] In seiner Orange Romania-Entscheidung legte der EuGH auch insoweit einen strengen Maßstab an und sah das Erfordernis, dass die Verweigerung der Einwilligung einer gesonderten schriftlichen Erklärung bedurfte, ohne weitere Prüfung des damit verbundenen Aufwands als geeignet an, die Freiwilligkeit der Entscheidung zu beeinflussen.[98] Dies spricht dafür, auch in anderen Konstellationen einen strengen Maßstab anzulegen, so z. B. wenn im Kontext von Cookie-Bannern der Weg zur Modifikation oder Ablehnung umständlicher als die Erteilung der Einwilligung. Hierfür spricht auch Vergleich mit dem Widerruf der

---

[94] *Weinzierl*, NVwZ Extra 2020, 1 (8).

[95] BeckOK DatenschutzR/*Stemmer*, DS-GVO Art. 7 Rn. 53.

[96] BeckOK DatenschutzR/*Stemmer*, DS-GVO Art. 7 Rn. 40.

[97] BeckOK DatenschutzR/*Stemmer*, DS-GVO Art. 7 Rn. 40: „Nachteil oder zusätzlicher Aufwand von einem gewissen Gewicht"; enger Gola/*Schulz*, DS-GVO Art. 7 Rn. 29: nur schwerwiegende Nachteile.

[98] EuGH NJW 2021, 841 Rn. 50 f. – Orange Romania.

Einwilligung: Muss dieser nach Art. 7 Abs. 3 S. 4 DSGVO so einfach wie die Erteilung der Einwilligung sein, so gilt dies im Erst-recht-Schluss auch für die primäre Verweigerung der Einwilligung.[99] Zu Recht wird hier sowohl von den Aufsichtsbehörden[100] als auch im Schrifttum[101] Äquivalenz gefordert.

Bereits nach der zum BDSG ergangenen Rechtsprechung des BGH konnte es an einer freien Entscheidung fehlen, wenn der Betroffene durch übermäßige Anreize finanzieller oder sonstiger Natur zur Preisgabe seiner Daten verleitet wurde.[102] Dies dürfte unter der DSGVO nach den autonom und einheitlich[103] auszulegenden Voraussetzungen für die Wirksamkeit einer Einwilligung weiterhin gelten. Unklar ist hingegen, wie bereits angesprochen, die Beurteilung bei einem die Entscheidung beeinflussenden Framing, durch das eine soziale Drucksituation erzeugt wird. Dies kann z. B. dadurch erfolgen, dass im Begleittext z. B. – zutreffend oder irreführend – auf die Knappheit eines Angebots hingewiesen wird (sog. Scarcity Patterns[104]), z. B. durch die Angabe der noch verfügbaren Exemplare oder durch von Countdowns begleitete, zeitliche befristete Angebote. Eine weitere denkbare Designoption besteht darin, dass mit der Erteilung der Einwilligung eine Steigerung bzw. mit ihrer Verweigerung eine Minderung des sozialen Status verbunden wird (sog. Emotional Steering[105] oder Social Proof Pattern[106]). Es spricht vieles dafür, dass auch ein solches Framing, wenn es für die angesprochenen Verkehrskreise hinreichend konkret und glaubhaft erscheint, im Einzelfall zu einer Mangel an Freiwilligkeit führen kann.[107]

c) Eindeutige und bestätigende Handlung

Eine Einwilligung in die Verarbeitung der eigenen Daten ist zudem nur dann wirksam, wenn sie auf einer eindeutigen, bestätigenden Handlung beruht (Art. 4 Nr. 11, Art. 6 Abs. 1 lit. a) DSGVO). Hieran fehlt es z. B. bei der Opt-Out-Ausgestaltung im Falle sog. Preselection-Patterns.[108]

---

[99] Ebenso schon *Sesing*, MMR 2021, 544 (547).

[100] *DSK*, Orientierungshilfe der Aufsichtsbehörden für Anbieter:innen von Telemedien, S. 13 f., 17.

[101] BeckOK DatenschutzR/*Stemmer*, 39. Ed. 1.11.2021, DS-GVO Art. 7 Rn. 40; *Sesing*, MMR 2021, 544 (547); *Lehr*/*Dietmann*/*Krisam*/*Volkamer*, DuD 2022, 296 (298).

[102] BGH NJW 2008, 3055 Rn. 21; BGH NJW 2010, 864 Rn. 21.

[103] EuGH MMR 2019, 732 Rn. 47 – Planet 49.

[104] Dazu *Martini*/*Kramme*/*Seeliger*, VuR 2022, 123 (123 f).

[105] *EDSA*, Guidelines 3/2022 on Dark patterns in social media platform interfaces, 2.

[106] Dazu s. *Luguri*/*Strahilevitz*, JLA 13 (2021), 43 (53).

[107] BeckOK DatenschutzR/*Stemmer*, DS-GVO Art. 7 Rn. 40; HK-DSGVO/*Ingold* Rn. 27; a. A. *Weinzierl*, NVwZ-Extra 2020, 1 (8).

[108] EuGH MMR 2019, 732 Rn. 60 ff. – Planet 49; BGH NJW 2020, 2540 Rn. 51 f. – Cookie-Einwilligung II; LG Rostock ZD 2021, 166 Rn. 49; *EDSA*, Guidelines 3/2022 on Dark patterns in social media platform interfaces Rn. 24.

2. Datenschutz durch Technikgestaltung

Gemäß Art. 25 Abs. 1 DSGVO hat der Verantwortliche geeignete technische und organisatorische Maßnahmen zu ergreifen, um Datenschutzgrundsätze wirksam umzusetzen. Dies gilt auch für den Grundsatz der Transparenz, der bei der Technikgestaltung, z. B. durch eine entsprechende Auswahl und Aufbereitung der zur Verfügung gestellten Informationen bei der Ausgestaltung der Benutzeroberfläche, zu berücksichtigen ist.[109] Allerdings ist fraglich, ob sich hieraus über die bisherige Darstellung hinausgehenden Transparenzvorgaben ergeben. Zumindest bislang bleiben die aus dieser Norm potentiell folgenden Vorgaben äußerst vage und bedürfen weiterer Konkretisierung.[110]

3. Zwischenergebnis

Auch wenn der Transparenzgrundsatz im Datenschutzrecht von fundamentaler Bedeutung ist, führt dessen Regelungsstruktur durch horizontale, allgemeinere Regelungen dazu, dass zentrale Fragen, insbesondere die an die Informationsverarbeitungskompetenz der betroffenen Personen anzulegenden Maßstäbe sowie die Auswirkungen eines emotionalen Framing auf die Freiwilligkeit einer Einwilligung, noch nicht hinreichend geklärt sind. Es ist zu hoffen, dass verschiedene Initiativen der Aufsichtsbehörden, insbesondere die vom EDSA zur öffentlichen Konsultation vorgelegten Leitlinien 3/2022 vom 14.3.2022, zu mehr Rechtssicherheit beitragen werden und dass auch der EuGH zeitnah Gelegenheit zur Stellungnahme erhalten wird.

IV. Fazit und gesetzgeberische Entwicklungen

Auf die Ausgestaltung von Nutzeroberflächen bezogene Transparenzpflichten sind zentrale Elemente für einen Interessenausgleich im Spannungsfeld von Verbraucherschutz und unternehmerischer Freiheit sowohl im Lauterkeits- als auch im Datenschutzrecht – „Transparency by Design" als Gegenmittel zu „Deception by Design". Allerdings ergeben sich bei ihrer Anwendung auf Dark Patterns zahlreiche ungeklärte Fragestellungen. Insofern ist es positiv zu vermerken, dass mit den Gesetzgebungsverfahren zum Digital Services Act sowie zum AI Act zwei gesetzgeberische Initiativen die Thematik von Dark Patterns oder zumindest subliminalen Beeinflussungen aufgreifen.

---

[109] S. die Beispiele bei *EDSA*, Leitlinien 4/2019 zu Artikel 25 vom 20.10.2020, Rn. 65 ff.

[110] Vgl. auch *Martini/Drews/Seeliger/Weinzierl*, ZfDR 2021, 47 (57 f); *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 56.

1. AI Act

Der von der Europäischen Kommission am 21.4.2021 vorgelegte Vorschlag für ein Gesetz über Künstliche Intelligenz enthält spezifische Transparenzpflichten für KI-Systeme. So enthält Art. 5 (1) a) ein allgemeines Verbot von KI-Systemen, die Techniken der unterschwelligen Beeinflussung einsetzen, wenn dies einen physischen oder psychischen Schaden verursachen kann. Hierunter dürften manipulative Designs fallen, die z. B. das Suchtpotential von Computerspielen erhöhen oder den Erwerb von gesundheitsgefährden Produkten einsetzen, auch wenn der Anwendungsbereich insgesamt auch aufgrund des weiten Begriffs der „Techniken der unterschwelligen Beeinflussung" unklar bleibt. Zudem haben Anbieter sicherzustellen, dass KI-Systeme, die für die Interaktion mit natürlichen Personen bestimmt sind, als solche erkennbar sind (Art. 52 Abs. 1). Weitere Transparenzpflichten gelten für Emotionserkennungssysteme (Art. 52 Abs. 2) und Deepfakes (Art. 52 Abs. 3).

2. Digital Services Act

Konkrete Vorschläge zur Regulierung von Dark Patterns im Online-Umfeld werden hingegen im Gesetzgebungsverfahrens des Digital Services Act diskutiert. Zwar hatte der Kommissions-Entwurf[111] Dark Patterns noch nicht adressiert. Sowohl der Rat der Europäischen Union als auch das Europäische Parlament haben nun entsprechende Vorschläge ergänzt.

Nach dem Vorschlag des Rates[112] soll Anbietern von Online-Marktplätzen untersagt werden, ihre Online-Schnittstelle so zu organisieren, dass die Nutzer getäuscht oder manipuliert werden sollen oder tatsächlich werden, indem ihre Autonomie, Entscheidungsfindung oder Wahlmöglichkeit untergraben oder beeinträchtigt wird. Zudem sollen sie dazu verpflichtet werden, ihre Online-Schnittstelle so zu organisieren, dass Unternehmer, die den Online-Marktplatz nutzen, ihren Informationspflichten hinsichtlich vorvertraglicher Informationen und Produktsicherheitsinformationen gemäß dem geltenden Unionsrecht nachkommen können. Schließlich sollen sie nach besten Kräften bewerten, ob die Unternehmer den unionsrechtlichen Verpflichtungen nachkommen (Art. 24b). Der Regelungstext erscheint wenig abgestimmt mit dem darüber hinaus gehenden Erw. 50a, der explizit Bezug auf „Dark Patterns" und die Transparenzvorgaben u. a. aus Art. 7 UGP-Richtlinie nimmt.

Der Vorschlag des Europäischen Parlaments[113] sieht in Art. 13a eine wesentlich detailliertere Regelung vor, die es Anbietern aller Arten von Vermittlungsdiens-

---

[111] *Europäische Kommission*, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie 2000/31/EG vor (COM/2020/825 final).

[112] *Europäischer Rat*, Allgemeine Ausrichtung vom 26.11.2021, 13203/21 ADD 1 REV 3.

[113] *Europäisches Parlament*, Position zum Digital Services Act vom 20.1.2022, P9_TA(2022)0014. Der in interinstitutionellen Verhandlungen vereinbarte Text (PE 734.311) wurde erst nach Einreichung des Manuskripts veröffentlicht.

ten, d. h. z. B. auch Sozialen Netzwerken, grundsätzlich untersagt, die Struktur, Funktion oder Funktionsweise ihrer Online-Schnittstelle dazu zu nutzen, um die Fähigkeit der Nutzer, eine freie, selbstständige und fundierte Entscheidung oder Wahl zu treffen, zu verzerren oder zu behindern, und einen nicht abschließenden Katalog verbotener Praktiken aufzählt. Hierbei nutzt das Parlament die Möglichkeit, um einige Fragestellungen explizit zu regeln, die durch die DSGVO nicht eindeutig beantwortet werden, z. B. die Unzulässigkeit der visuellen Hervorhebung der Einwilligungsoption. Zudem soll die Kommission dazu ermächtigt werden, die Liste unzulässiger Praktiken durch delegierte Rechtsakte zu ergänzen.

Unter dem Gesichtspunkt der Rechtssicherheit erscheint es grundsätzlich sinnvoll, wenn durch den DSA offene, konkrete Fragestellungen geklärt werden können,[114] da eine Anpassung der DSGVO selbst derzeit unrealistisch erscheint. Die Zulässigkeit bestimmter Praktiken wird sich hingegen häufig nur anhand der Umstände des Einzelfalls feststellen lassen, so dass es neben den konkreten Vorgaben grundsätzlich auch offener, wertungsabhängiger Tatbestände bedarf. Allerdings erschließt sich auf den ersten Blick nicht ohne weiteres, welchen Mehrwert ein offener Tatbestand wie z. B. das allgemeine Täuschungs- und Manipulationsverbot in Art. 24b Abs. -1 des Vorschlags des Rates gegenüber den sich z. B. aus Art. 7 UGP-Richtlinie ergebenden Vorgaben haben würde, zumal die vorgeschlagene Regelung aufgrund ihrer hinter der UGP-Richtlinie zurückbleibenden Detailtiefe (z. B. hinsichtlich der an die Nutzer hinsichtlich Informationsverarbeitungskompetenzen anzulegenden Maßstäbe) für ein erhebliches Maß an Rechtsunsicherheit sorgen dürfte.[115] Insofern wäre eine genauere Bestimmung des Verhältnisses zwischen dem DSA und den bestehenden Regelungen wünschenswert.

[114] A. A. *Kühling/Sauerborn*, Rechtliche Rahmenbedingungen sogenannter „Dark Patterns", 60.
[115] *Martini/Kramme/Seeliger*, VuR 2022, 123, 129 f.

Rechtsschutz gegen diskriminierende „KI“

Caroline Meller-Hannich / Lukas Hundertmark

I. Einleitung

Der zunehmende und inzwischen verbreitete Einsatz von Algorithmen bzw. „Künstlicher Intelligenz (KI)“[1] durch private Akteure[2] im Rechts- und Geschäftsverkehr birgt die Gefahr, dass es zu Diskriminierungen kommt – aufgrund des Geschlechts, der Herkunft, der sexuellen Orientierung, also auch und gerade in grund- und menschenrechtlich geschützten Bereichen der menschlichen Existenz. Die Europäische Kommission hat auf diese Entwicklungen kürzlich mit dem Vorschlag für eine KI-Verordnung (KI-VO-E)[3] reagiert.

Ebenso wichtig wie die materiell-rechtliche Regulierung des Einsatzes diskriminierender Algorithmen sind freilich die Durchsetzungsmöglichkeiten bestehender Rechte. Da Algorithmen nicht für den einmaligen Einsatz konzipiert sind, kommt es bei Diskriminierungen typischerweise gleich zu einer Gefährdung einer Vielzahl von Personen. Deshalb ist es wichtig, nicht nur individuelle Rechtsdurchsetzungsmöglichkeiten in den Blick zu nehmen, sondern zu überlegen, inwiefern Mittel des kollektiven Rechtsschutzes – insbesondere Verbandsklagen – gegen diskriminierende Algorithmen eingesetzt werden können.

---

[1] Die im Beitrag thematisierten Probleme treten allgemein bei der Verwendung von Algorithmen auf, sodass vorrangig dieser Begriff und nicht derjenige der „KI“ benutzt wird. Sofern dennoch von „KI“ oder „KI-Systemen“ gesprochen wird, geschieht das, um die Terminologie des KI-VO-E zu gebrauchen und ist als Synonym zu Algorithmen zu verstehen; zu den Schwierigkeiten der Verwendung des Begriffs „KI“, insbesondere in der Rechtswissenschaft s. nur *Herberger* NJW 2018, 2825; *Timmermann*, Legal Tech-Anwendungen, 2020, 60 ff.

[2] m/w/d.

[3] Vorschlag für eine Verordnung zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, COM (2021) 206 final.

II. Diskriminierende Algorithmen

1. Anwendungsfälle diskriminierender Algorithmen

Beim Einsatz von Algorithmen im privatrechtlichen Bereich kann es zu verschiedenen Formen von Diskriminierung kommen. Da Algorithmen häufig zur Entscheidungsfindung im Vorfeld oder bei Abschluss eines Vertrags eingesetzt werden, kann eine Diskriminierung durchaus handfeste vermögensrechtliche Folgen haben und sogar so weit reichen, dass Personen vom Zugang zu bestimmten Gütern und Dienstleistungen ausgeschlossen werden. Denkbare Probleme sollen folgende Fälle illustrieren[4]:

Fall 1: X erhält bei einem Mobilfunkanbieter keinen Handyvertrag – der Sachbearbeiter sagt, das habe „die KI" so entschieden, er selbst könne sich das nicht erklären. Er werde dem Vorschlag „der KI" folgen und rät zum Abschluss des Vertrages über eine Freundin, für die „die KI" einen Vertrag anbietet.

Fall 2: X berät sich mit ihrer Bank über einen Immobilienkredit zum Zwecke des Erwerbs einer Eigentumswohnung, ein erstes Angebot wird kurz vor Vertragsabschluss durch „die KI" überprüft – diese zeigt an, dass der Kredit an X als alleinerziehende Mutter nicht vergeben werden könne; von der „KI" abweichende Entscheidungen können nicht getroffen werden.

Algorithmen können ermitteln, ob eine Person bestimmte Kriterien für einen Vertragsschluss erfüllt (*scoring* bzw. *rating*). Neben der erwähnten Situation der Kreditvergabe oder des Vertragsschlusses über Güter oder Dienstleistungen kann das z. B. auch bei Auswahlgesprächen im Bewerbungsprozess um eine Arbeitsstelle der Fall sein.[5] Algorithmen können dabei auch personenbezogene Daten auswerten und zusammenführen, um das Gesamtbild einer Persönlichkeit zu bewerten (*profiling*, Art. 4 Nr. 4 Datenschutz-Grundverordnung (DSGVO)). Diese Datenanalysen können Unternehmer etwa nutzen, um bestimmten Personen(gruppen) einen Vertragsschluss zu verweigern oder Vertragsbedingungen für sie anzupassen (Preisdifferenzierung). Algorithmen können hier diskriminierend wirken – entweder, weil sie explizit darauf programmiert sind, bestimmte persönliche Merkmale negativ zu gewichten, oder weil sie Vergleichsdaten verwenden, denen implizit bereits eine Diskriminierung zugrunde liegt. Dabei besteht die Gefahr, dass Algorithmen aus bloßer Korrelation von Daten auf kausale Zusammenhänge schließen, etwa anhand von Wohnort oder Namen einer Person auf deren Kreditwürdigkeit schließen.[6] Auch wenn Algorithmen auf den ersten Blick Objektivität

---

[4] Für die Praxisbeispiele ist unserer Mitarbeiterin und Kollegin Dr. Katharina Gelbrich zu danken; weitere Beispiele bei *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, 34 ff.

[5] Zum Einsatz von Algorithmen im Bewerbungsverfahren *Joos* NZA 2020, 1216; *Hoffmann* NZA 2022, 19; zur Diskriminierung dabei *Dzida*/*Groh* NJW 2018, 1917; *Steege* MMR 2019, 715, 718 f.; *Grünberger* ZRP 2021, 232.

[6] Vgl. *Martini* JZ 2017, 1017, 1018.

versprechen, sind ihre Entscheidungen abhängig von ihrer Programmierung und der Qualität der bereits vorhandenen Daten. Auf diese Weise perpetuieren sich (möglicherweise unbewusste) Vorurteile und vorhandene strukturelle Probleme.[7] Für Angehörige bestimmter Personengruppen besteht dabei das Risiko, pauschal von Teilen des Geschäfts- und Privatrechtsverkehrs ausgeschlossen zu werden.

Vielfach werden Algorithmen unmittelbar in die Entscheidungsfindung einbezogen. Dabei ist zwischen sog. letztentscheidenden und lediglich unterstützenden Algorithmen zu unterscheiden. Erstere treffen bereits eine endgültige Entscheidung (so in unserem Beispielsfall 2), während letztere eine Entscheidungsgrundlage für eine menschliche Entscheidung schaffen, etwa indem sie eine Prognose anbieten oder einen konkreten Vorschlag liefern (so in unserem Beispielsfall 1). Besonders letztentscheidende Algorithmen bergen Risiken, da Menschen in der Regel im Einzelfall differenziertere Wertungen vornehmen können. So wird in Beispielsfall 2 die Person X aufgrund eines persönlichen Merkmals (alleinerziehende Mutter) von der Kreditvergabe durch den Algorithmus ausgeschlossen. Gegenüber einem menschlichen Entscheider könnte X noch zusätzliche Argumente vorbringen, warum eine Rückzahlung des Darlehens entgegen der Prognose des Algorithmus, zu erwarten ist, sodass es möglicherweise noch zu einer anderen Entscheidung kommen kann. Diese Möglichkeit besteht bei letztentscheidenden Algorithmen nicht. Nicht zuletzt werden viele Personen eine menschliche Entscheidung anders (mutmaßlich verständnisvoller) aufnehmen als eine automatisch getroffene Entscheidung.

Dies wird meist noch durch die Intransparenz von Algorithmen verschärft. Oft kann nicht ohne weiteres nachvollzogen werden, warum sie ein bestimmtes Ergebnis liefern. Das wird in Fall 1 deutlich, in dem nicht nachvollziehbar ist, warum der Algorithmus sich gegen einen Vertragsschluss ausspricht. Intransparenz ist insbesondere für solche Algorithmen typisch, die mittels *machine learning* arbeiten und auf diese Weise zunehmend zu unvorhergesehen, für Menschen nicht nachvollziehbaren Ergebnissen gelangen.[8] Das macht es einerseits schwierig, festzustellen ob die Algorithmen diskriminierend wirken[9] (dazu näher unter IV.2.). Andererseits erschwert es auch, korrigierend einzugreifen. Wenn ein Mensch weder den Weg noch das Ergebnis des Einsatzes eines Algorithmus´ nachvollziehen kann, wird er schwerlich geeignete Anknüpfungspunkte für eine Korrektur finden.

2. Ausreichender Schutz gegen diskriminierende Algorithmen durch das materielle Recht?

Der diskriminierende Einsatz von Algorithmen wird bislang nur teilweise von den im nationalen und europäischen Recht vorhandenen materiell-rechtlichen Rege-

---

[7] Vgl. *Ernst* JZ 2017, 1026, 1028; *Martini* (Fn. 6), 1018 f.; *Sesing/Tschech* MMR 2022, 24.
[8] *Grünberger* (Fn. 5), 233; *Ebers*, in: Ebers/Navas (Hrsg.), Algorithms and law, 2020, 47 ff.
[9] *Hoffmann-Riem* AöR 142 (2017), 1, 32 f.; *Martini* (Fn. 6), 1018.

lungen erfasst. Die möglichen Anwendungsfälle berühren typischerweise nicht einzelne Rechtsgebiete, sondern spielen sich in vielen verschiedenen Bereichen des materiellen Rechts ab. Die oben genannten Fragen können z. B. Relevanz für das Vertragsrecht, Deliktsrecht, Produkthaftungsrecht, Datenschutzrecht, Lauterkeitsrecht, Kartellrecht oder Arbeitsrecht haben.

Dementsprechend ergeben sich die Ansprüche, die gegen den Einsatz diskriminierender Algorithmen in Betracht kommen, aus vielfältigen Gesetzen. So gewährt § 1 Abs. 1 S. 1 Produkthaftungsgesetz (ProdHaftG) gegenüber dem Hersteller eines fehlerhaften Produkts einen Anspruch auf Ersatz von Schäden an Leben, Körper oder Gesundheit, die durch das fehlerhafte Produkt entstehen. Bei diskriminierenden Algorithmen könnte es sich um solche fehlerhaften Produkte handeln. Nach § 8 Abs. 1 S. 1 des Gesetzes gegen den unlauteren Wettbewerb (UWG) kann derjenige, der eine unzulässige geschäftliche Handlung vornimmt, auf Beseitigung oder Unterlassung in Anspruch genommen werden. Dies kommt z. B. für die Verwendung von diskriminierenden Algorithmen in Betracht, soweit diese durch andere Rechtsvorschriften untersagt ist. Soweit dies zu einer Wettbewerbsbeschränkung führt, ist auch das Gesetz gegen Wettbewerbsbeschränkungen (GWB) heranzuziehen.[10] Hier kommen ebenfalls Ansprüche auf Beseitigung und Unterlassen (§ 33 Abs. 1 GWB) oder Schadenersatz (§ 33a Abs. 1 GWB) in Betracht.

Auch das Allgemeine Gleichstellungsgesetz (AGG) hält in §§ 15 und 21 AGG Beseitigungs- und Ersatzansprüche bei bestimmten Benachteiligungen (etwa wegen des Geschlechts, der ethnischen Herkunft oder der sexuellen Identität) bereit. Davon erfasst sein könnten z. B. Arbeitgeber, die im Bewerbungsprozess diskriminierende Algorithmen verwenden (§ 15 Abs. 1 S. 1 AGG).[11] Schließlich kommen auch deliktische Ansprüche nach § 823 Abs. 1 Bürgerliches Gesetzbuch (BGB) in Frage, z. B. wenn diskriminierende Algorithmen Körper, Gesundheit oder allgemeines Persönlichkeitsrecht verletzen.

Meist ergeben sich die materiellen Rechte damit aus Normen, die zwar keinen spezifischen Bezug zu Algorithmen aufweisen, aufgrund ihrer meist technikneutralen Konzeption aber ohne weiteres auch auf den diskriminierenden Einsatz von Algorithmen anwendbar sind. Speziell auf Rechtsverletzungen durch Algorithmen bezogene Ansprüche existieren bisher nur wenige, wenngleich immer wieder Forderungen für einen „KI"-spezifischen Haftungstatbestand auftauchen.[12]

Eine Ausnahme bietet derzeit etwa Art. 22 DSGVO, der dem Einsatz von Algorithmen zur Entscheidungsfindung enge Grenzen zu setzen scheint. Die Vorschrift betrifft nur solche Fälle, in denen eine Entscheidung ausschließlich automatisiert erfolgt. Damit ist nur die Anwendung letztentscheidender Algorithmen untersagt. Algorithmen, die eine Entscheidung vorschlagen, die dann ein Mensch umsetzt

---

[10] S. für Preisdiskriminierung durch Algorithmen *Paal* GRUR 2019, 43.
[11] S. dazu *Sesing/Tschech* (Fn. 7), 25 ff.
[12] *Martini* (Fn. 6), 1024.

(z. B. das *scoring*), sind nicht erfasst.[13] Allerdings scheint zumindest fraglich, inwiefern der menschliche Entscheider das Ergebnis in Fällen hinterfragen wird, in denen der Algorithmus ihm die Arbeit bereits erspart hat und er die einzelnen Arbeitsschritte des Algorithmus ohnehin nicht nachvollziehen kann.[14] Im Ergebnis macht es für betroffene Personen insofern einen geringen Unterschied, ob ein Algorithmus letztentscheidet oder ein Mensch den Entscheidungsvorschlag ungeprüft übernimmt. So wäre in Fall 2 die Datenverarbeitung des letztentscheidenden Algorithmus unzulässig. Die Konstellation in Fall 1 wäre allerdings zulässig, da der Sachbearbeiter die theoretische Möglichkeit zur anderen Entscheidung hatte, auch wenn er sie nicht wahrgenommen hat.[15] Für X dürfte diese Feinheit nur wenig tröstlich sein. Auch für letztentscheidende Algorithmen können im Übrigen die weitreichenden Ausnahmen des Art. 22 Abs. 2 DSGVO greifen. Falls etwa die ausschließlich automatische Verarbeitung für Abschluss oder Erfüllung eines Vertrages erforderlich ist oder sie mit Einwilligung des Betroffenen erfolgt, ist die Verarbeitung zulässig.

Es gibt weitere spezifische Regelungen für Algorithmen, die sich allerdings in Hinweis- und Informationspflichten erschöpfen, beispielsweise der zum 1.10.2021 in Kraft getretene § 13b Abs. 2 Rechtsdienstleistungsgesetz (RDG). Demnach müssen Verbraucher, deren Mandat von einem Inkassodienstleister abgelehnt wird, darüber informiert werden, ob eine automatisierte rechtliche Prüfung der Forderung stattgefunden hat. Durchsetzbare Ansprüche für die betroffenen Verbraucher ergeben sich daraus jedoch noch nicht, da es sich bei den Informationspflichten um nicht einklagbare Nebenpflichten handelt. Ein Verstoß gegen diese Pflichten kann zwar Anknüpfungspunkt für einen Schadenersatzanspruch sein, allerdings dürfte der Nachweis eines Schadens regelmäßig schwerfallen.

Insgesamt decken die allgemeinen materiell-rechtlichen Regelungen zwar viele Probleme diskriminierender Algorithmen ab. Der Schutz ist dennoch nur fragmentarisch. Insbesondere bei hochentwickelten KI-Anwendungen, die schnell zu unvorhergesehenen und bisher unbekannten Problemstellungen führen können, ist dies riskant.[16] Das materielle Recht ist aus diesen Gründen derzeit nicht in vollem Umfang in der Lage, die mit dem Einsatz diskriminierender Algorithmen einhergehenden rechtlichen Probleme einer umfassenden sachgerechten Lösung zuzuführen.

---

[13] *Ebers* (Fn. 8), 52; *Martini* (Fn. 6), 1020.

[14] *Hoffmann-Riem* (Fn. 9), 35 f.; *Ebers* (Fn. 8), 52.

[15] Das VG Wiesbaden ZD 2022, 121 m. Anm. *Qasim* sieht dagegen in einer Vorlage an den EuGH auch das automatische Erstellen der Score-Werte als „Entscheidung“ i. S. d. Art. 22 Abs. 1 DSGVO an.

[16] Vgl. *Martini* (Fn. 6), 1021.

3. Sektorspezifische Regelungen im KI-VO-E

Mit dem KI-VO-E liegt nunmehr ein ganzheitlicher Regulierungsvorschlag auf europäischer Ebene vor. Der KI-VO-E reguliert KI-Systeme. Wann es sich bei Software um ein KI-System handelt, ist dabei nach Art. 3 Nr. 1 KI-VO-E i. V. m. Anhang I abhängig von der zur Entwicklung verwendeten Technik. Dies betrifft Konzepte des maschinellen Lernens ebenso wie logikgestützte Konzepte oder statistische Ansätze. Der Begriff der KI-Systeme ist damit derart weit gefasst, dass wohl nur sehr einfache Softwareanwendungen vom Anwendungsbereich des Verordnungsentwurfs ausgenommen sind.[17] Insofern will die Kommission auch Software regulieren, von der keine „KI"-spezifischen Risiken ausgehen.

Der KI-VO-E verfolgt einen risikobasierten Ansatz und unterscheidet zwischen KI-Systemen mit minimalem, geringem, hohen oder unannehmbaren Risiko, an deren Betrieb entsprechend unterschiedliche Anforderungen gestellt werden.

Art von KI-System	*Beispiel*	*Rechtsfolge*
KI-System mit unannehmbarem Risiko	Manipulative, schädigende Beeinflussung	Verbot
KI-System mit hohem Risiko	Bewertung von Kreditwürdigkeit	Kontrolle
KI-System mit geringem Risiko	Chatbots	Informationspflicht
Kein KI-System (minimales Risiko)	Spamfilter	Nicht vom KI-VO-E erfasst

Tabelle 1: risikobasierte Einteilung von KI-System nach dem KI-VO-E.

Für KI-Systeme mit unannehmbarem Risiko statuiert Art. 5 KI-VO-E ein generelles Verbot. Hierunter fallen etwa KI-Systeme, die Personen derart beeinflussen, dass dies zu physischen oder psychischen Schäden führt. Auch das *social scoring*, also die Bewertung der Vertrauenswürdigkeit einer natürlichen Person aufgrund ihres sozialen Verhaltens oder bestimmter persönlicher Merkmale, ist hiernach verboten. Dieses Verbot bezieht sich allerdings nur auf den behördlichen Einsatz, nicht den privatwirtschaftlichen Bereich.

Für die Entwicklung, Beaufsichtigung und Verwendung von Hochrisiko-KI-Systemen werden in Art. 16–29 KI-VO-E umfangreiche Pflichten bestimmt.[18] Die Pflichten bauen aufeinander auf und richten sich an Anbieter, Importeur, Händler und Nutzer solcher Systeme.[19] Zudem müssen die Systeme eine Konformitätsbewertung durchlaufen, bevor sie in den Verkehr gebracht oder in Betrieb genom-

[17] *Roos/Weitz* MMR 2021, 844, 845; *Bomhard/Merkle* RDi 2021, 276, 277; *Ebers/Hoch/Rosenkranz/Ruschemeier/Steinrötter* RDi 2021, 538, 539.

[18] S. dazu *Geminn* ZD 2021, 354, 357 f.; krit. zur Ausgestaltung *Ebers/Hoch/Rosenkranz/Ruschemeier/Steinrötter* (Fn. 15), 543 f.

[19] S. zu dazu *Roos/Weitz* (Fn. 15), 846 ff.

men werden (Art. 19, 43 KI-VO-E). Hierunter fallen z. B. Sicherheitskomponenten von Produkten (Art. 6 Abs. 1 KI-VO-E) sowie KI-Systeme, die in bestimmten besonders sensiblen Bereich eingesetzt werden (Art. 6 Abs. 2 KI-VO-E i. V. m. Anhang III). Letzteres betrifft etwa die Identifizierung und Kategorisierung von Personen (Anhang III Nr. 1) oder Bewerbungs- und Einstellungsprozesse (Anhang III Nr. 4). Die in unserem Beispielsfällen eingesetzten Algorithmen sind demnach auch als Hochrisiko-KI-Systeme anzusehen, da sie für die Kreditwürdigkeitsprüfung und Kreditpunktebewertung natürlicher Personen verwendet werden (Anhang III Nr. 5 lit. b).

Für Anwendungen mit geringem Risiko sind hingegen nur Informationspflichten vorgesehen, die bei der Interaktion mit natürlichen Personen transparent machen, dass man es mit einem KI-System zu tun hat (Art. 52 KI-VO-E). Hierunter dürften z. B. regelmäßig Chatbots fallen.

Für die übrigen Anwendungen wird nach dem KI-VO-E ein minimales Risiko angenommen, welches keine zusätzlichen Anforderungen erforderlich macht. Dabei dürfte es sich um einfachste Anwendungen, z. B. Spamfilter, handeln.

Der KI-VO-E sieht bei Verstößen gegen die beschriebenen Vorgaben und Pflichten für die davon betroffenen Personen keine eigenen materiellen Rechte vor. Fragen der Haftung werden (wie auch sonstige zivilrechtliche Fragestellungen) ausgeklammert. Statt auf privatrechtliche Rechtsdurchsetzung setzt der Entwurf auf Prävention durch Zertifizierungsverfahren und die Androhung von Bußgeldern (Art. 71, 72 KI-VO-E). Ob dies ausreichend ist, ist allerdings zu bezweifeln.[20] Wer in seiner Selbstbestimmung, seinem Eigentum, Vermögen und/oder Marktzugang durch diskriminierende KI gehindert oder verletzt ist, sollte eigene materiell-rechtliche Abhilfeansprüche haben, um Unterlassung oder ggf. sogar Schadenersatz geltend machen zu können.

4. Zwischenergebnis

Für die vielfältigen Anwendungsfälle diskriminierender Algorithmen hält das materielle Recht nur selten technikspezifische Regelungen bereit. Über allgemein gefasste Anspruchsgrundlagen, z. B. § 1 ProdHaftG, § 823 Abs. 1 BGB, § 8 Abs. 1 S. 1 UWG, §§ 33 Abs. 1, 33a Abs. 1 GWB, §§ 15 bzw. 21 AGG bestehen durchaus individuelle Rechtsschutzmöglichkeiten gegen diskriminierende Algorithmen. Die Wirksamkeit der materiell-rechtlichen Regeln hängt jedoch entscheidend von ihrer praktischen Durchsetzung ab. Dabei stößt die individuelle Rechtsdurchsetzung meist dann an ihre Grenzen, wenn die Rechtsverletzung den Einzelnen nur in geringem Maß betrifft, und Aufwand und Risiko der Rechtsdurchsetzung den Nutzen übersteigen.[21] In diesen Fällen besteht ein rationales Desinteresse

[20] Krit. auch *Ebers/Hoch/Rosenkranz/Ruschemeier/Steinrötter* (Fn. 15), 545 f.

[21] S. nur *Meller-Hannich*, Gutachten A zum 72. Deutschen Juristentag, 2018, A 24 f.

an der Rechtsdurchsetzung, was regelmäßig zum Verzicht auf diese führen wird. Dies dürfte in Fällen diskriminierender Algorithmen durchaus der Fall sein. Der KI-VO-E reguliert zwar umfangreich Herstellung, Inverkehrbringen und Verwendung von KI-Systemen, sieht aber keine Möglichkeiten der privatrechtlichen Rechtsdurchsetzung vor.

III. Kollektive Rechtsschutzmöglichkeiten

Betreffen Rechtsverletzungen eine Vielzahl von Fällen, besteht ein objektives Interesse daran, dass gegen die Verletzungen vorgegangen wird. Möglichkeiten dazu bietet der kollektive Rechtsschutz.

Algorithmen können dabei übrigens nicht nur Gegenstand, sondern auch Hilfsmittel des kollektiven Rechtsschutzes sein. Algorithmen werden schon regelmäßig eingesetzt, um Ansprüche gebündelt geltend zu machen, etwa von Legal Tech Anbietern, und bieten noch weiteres Potenzial im Rahmen von Verbandsklagen, etwa in Form von elektronisch geführten Klageregistern. Algorithmen selbst könnten damit den Rechtsschutz gegen diskriminierende Algorithmen unterstützen.[22]

1. Verbandsklagen nach dem UKlaG und dem UWG

In Deutschland existiert bereits eine Reihe von Verbandsklagen, die als Mittel des kollektiven Rechtsschutzes gegen diskriminierende Algorithmen in Betracht kommen.[23] Verbraucherschutzverbände können (als qualifizierte Einrichtungen i. S. d. § 3 Abs. 1 S. 1 Nr. 1 (Unterlassungsklagengesetz (UKlaG) bzw. § 8 Abs. 3 Nr. 3 UWG) selbst Ansprüche nach dem UKlaG bzw. dem UWG durchsetzen. Während das UKlaG Schutz gegen die Verwendung rechtswidriger AGB (§ 1 UKlaG) und gegen verbraucherschutzgesetzwidrige Praktiken (§ 2 UKlaG) bieten soll, richtet sich das UWG gegen unlautere geschäftliche Handlungen, dient also zuvörderst dem fairen Wettbewerb im Anbieter- und Verbraucherinteresse.

Im Bereich des UKlaG scheinen nur Ansprüche nach § 2 UKlaG ernsthaft gegen diskriminierende Algorithmen in Betracht zu kommen. Insbesondere sind nach § 2 Abs. 2 S. 1 Nr. 11 UKlaG solche Vorschriften als Verbraucherschutzgesetze erfasst, welche die Zulässigkeit der Erhebung, Verarbeitung oder Nutzung personenbezogener Daten regeln. § 2 UKlaG ist allerdings in der Anwendung insoweit unscharf, als zunächst immer festgestellt werden muss, ob eine bestimmte Vorschrift verbraucherschützenden Charakter hat. Dazu genügt es nicht, dass eine

---

[22] S. Beitrag *Rühl*, in diesem Band S. 269.

[23] S. allgemein zu kollektiven Rechtsschutzmöglichkeiten bei Diskriminierung *Herberger* RdA 2022, 220, 221 ff.

Vorschrift (auch) verbraucherschützende Wirkung zeigt, sie muss vielmehr Personen gerade in ihrer Eigenschaft als Verbraucher schützen.[24] Aus diesem Grund ist etwa § 19 AGG, der die zivilrechtliche Diskriminierung aufgrund ethnischer Herkunft oder sexueller Identität verbietet, kein Verbraucherschutzgesetz. Die Norm zielt allgemein auf den Schutz natürlicher Personen, nicht aber speziell von Verbrauchern ab.[25] Auch ist nach § 15 UKlaG das Arbeitsrecht pauschal vom Anwendungsbereich der Unterlassungsklage nach dem UKlaG ausgenommen. Gleiches dürfte für die Normen des KI-VO-E gelten, dessen Vorschriften und Erwägungsgründen man keinen spezifischen Verbraucherbezug entnehmen kann. Der KI-VO-E soll vielmehr alle Personen und den Binnenmarkt als solchen schützen.

Eine Verbandsklage nach dem UWG gegen den Einsatz von KI-Systemen ist möglich, wenn dieser Einsatz gegen wettbewerbsrechtliche Vorschriften verstößt. Soweit ein Verstoß gegen die KI-Verordnung eine unlautere geschäftliche Handlung i. S. d. § 3 UWG darstellt oder den Tatbestand des Rechtsbruchs, § 3a UWG, erfüllt, ist eine Klagemöglichkeit (auch) für Verbände eröffnet. § 8 Abs. 1 S. 1 UWG begrenzt die Rechtsfolgen allerdings auf Beseitigung und Unterlassen der unlauteren Handlung. Zusätzlich kommt eine Klage auf Gewinnabschöpfung gemäß § 10 UWG in Betracht, wobei diese sie an hohe Hürden geknüpft und für die klagenden Verbände regelmäßig unattraktiv ist.[26]

Schadenersatz nach § 9 UWG können Verbände nicht geltend machen, da dieser Anspruch nur Mitbewerbern zusteht. Nicht zu unterschätzen sein dürfte dabei der Anreiz der Mitbewerber, gegen wettbewerbswidrigen Einsatz von Algorithmen durch die Konkurrenz vorzugehen – dabei handelt es sich freilich nicht um Verbandsklagen und Verbände können darauf auch nicht unmittelbar Einfluss nehmen.

Seit Neuestem können nach § 9 Abs. 2 UWG auch Verbraucher Schadenersatz verlangen, wenn sie durch unzulässige geschäftliche Handlungen zu geschäftlichen Entscheidungen veranlasst werden, die sie andernfalls nicht getroffen hätten. Ob dies zu einer gesteigerten Rechtsdurchsetzung durch die Verbraucher selbst führen wird, scheint fraglich, da das rationale Desinteresse im Bereich geringfügiger Forderungen bestehen bleiben wird. Allerdings kann der Anspruch einen Anknüpfungspunkt für die neue Verbandsklage bieten, mit der die Verbraucheransprüche gesammelt durchgesetzt werden können (s. dazu unten IV.).

Vollumfänglichen Rechtsschutz gegen diskriminierende KI bieten die Verbandsklagen nach UKlaG und UWG damit insgesamt nicht.[27]

---

[24] Köhler/Bornkamm/Feddersen/*Köhler*, UWG, 39. Auflage 2021, § 2 UKlaG, Rn. 2.

[25] OLG Hamm, Urteil vom 3.3.2017 – 12 U 104/16; LG Kiel, Urteil vom 28.5.2015 – 17 O 79/15; offengelassen von OLG Schleswig-Holstein VuR 2016, 190, 191.

[26] Vgl. Fezer/Büscher/Obergfell/*von Braunmühl*, Lauterkeitsrecht: UWG, 3. Auflage 2016, § 10 UWG, Rn. 152 ff.

[27] S. speziell für Durchsetzung des AGG *Braunroth*, VuR 2018, 455, 457.

2. Musterfeststellungsklagen

Seit dem 1.11.2018 haben qualifizierte Einrichtungen die Möglichkeit, Musterfeststellungsklagen nach den §§ 606 ff. Zivilprozessordnung (ZPO) zur Feststellung der tatsächlichen und rechtlichen Voraussetzungen von Ansprüchen oder Rechtsverhältnissen zwischen Verbrauchern und Unternehmern zu erheben. Verbraucher haben die Möglichkeit, sich über ein Klageregister zur Klage anzumelden und sind dann an die Wirkungen eines Musterfeststellungsurteils gebunden.

In ihrem Anwendungsbereich kommt die Musterfeststellungsklage theoretisch durchaus als Rechtsschutzinstrument gegen diskriminierende KI in Betracht. Voraussetzung dafür ist allerdings, dass das materielle Recht entsprechende Ansprüche von Verbrauchern gegen Unternehmer bereithält. Lediglich auf Streitigkeiten, die der Arbeitsgerichtsbarkeit unterfallen, findet die Musterfeststellungsklage wiederum nach § 46 Abs. 2 S. 2 Arbeitsgerichtsgesetz (ArbGG) keine Anwendung. Aber auch zur Durchsetzung der Rechte aus dem AGG scheint die Musterfeststellungsklage aufgrund der Anforderungen an qualifizierte Einrichtungen nicht geeignet.[28]

Problematisch ist zudem, dass Verbraucher zweifach aktiv werden müssen. Zum einen für die Anmeldung zum Klageregister, zum anderen müssen sie nach dem Musterfeststellungsverfahren noch ein Individualverfahren betreiben, um einen Leistungstitel zu erhalten, was für den Verbraucher zusätzliche Unsicherheit bedeutet und die Effektivität der Musterfeststellungsklage abschwächt.[29]

Nach fast vierjährigem Bestehen wurden lediglich 29 Musterfeststellungsklagen erhoben – eine gegen Diskriminierung gerichtete war nicht darunter.[30] Die Musterfeststellungsklage bietet damit ebenfalls keinen hinreichenden kollektiven Rechtsschutz gegen diskriminierende Algorithmen.

3. Spezielle Verbandsklagerechte von Antidiskriminierungsverbänden

Schließlich werden Antidiskriminierungsverbänden mitunter spezielle Rechte eingeräumt. So haben sie etwa die Möglichkeit, nach § 15 Behindertengleichstellungsgesetz (BGG)[31] Verstöße gegen Vorschriften zur Gleichstellung behinderter Menschen feststellen zu lassen. Auch das Berliner Landesantidiskriminierungsgesetz (LADG) sieht in § 9 eine Möglichkeit speziell für Antidiskriminierungsverbände vor, auf Feststellung von Verstößen gegen das Diskriminierungsverbot zu klagen.

Im Übrigen können Antidiskriminierungsverbände benachteiligte Personen bei der Durchsetzung ihrer Rechte in gerichtlichen Verfahren unterstützen. Der

---

[28] Dazu *Braunroth* (Fn. 24), 458 ff.

[29] S. nur *Meller-Hannich* (Fn. 21), A 47.

[30] S. die Bekanntmachungen im Klageregister des Bundesamts für Justiz, abrufbar unter https://www.bundesjustizamt.de/DE/Themen/Buergerdienste/Klageregister/Bekanntmachungen/Klagen_node.html;jsessionid=CE2E36DB4657AF80ED15180DA4333899.1_cid501 (Stand: 2.9.2022).

[31] S. auch entsprechende landesrechtliche Regelungen, etwa § 19 BGG LSA, Art. 17 BayBGG, § 17 HesBGG.

dafür einschlägige § 23 AGG gewährt den Verbänden jedoch kein eigenes Klagerecht, sodass es sich hier nicht um echte Verbandsklagen handelt. Es geht dabei eher darum, individuell betroffene Personen in ihrem eigenen Prozess zu unterstützen, weniger um kollektiven Rechtsschutz.[32]

In diesem Bereich wird deshalb bereits seit längerem in der Rechtswissenschaft über eine Ausweitung der Rechtsschutzmöglichkeiten gegen Diskriminierung diskutiert. Zum einen wird dafür die Einführung einer spezifischen, auf die Prävention von Diskriminierung gerichteten Verbandsklage erwogen.[33] Zum anderen wird die Ausweitung der Ersatzansprüche aus §§ 15 und 21 AGG gefordert.[34] Das bisherige Bild einer schwer zugänglichen oder gar nicht vorhandenen kollektiven Rechtsschutzlandschaft gegen den diskriminierenden Einsatz von Algorithmen spricht dafür, diese Ansätze weiter zu verfolgen.

4. Zwischenergebnis

Insgesamt bestehen somit durchaus Möglichkeiten, sich mittels Verbandsklagen gegen diskriminierende Algorithmen zu wehren, wenngleich der Schutz auf diese Weise nicht umfassend möglich ist. Dabei bestehen keine KI-spezifischen Verbandsklagerechte, sondern es wird an die allgemeinen materiell-rechtlichen Klagemöglichkeiten angeknüpft. Eine Ausweitung der Klagebefugnisse und die Möglichkeit der Verbände, unmittelbar Leistungen für die Verbraucher einzuklagen, könnten zu einem effektiven Rechtsschutz beitragen. Ehrlicherweise muss aber auch gesagt werden: In den Bereichen, in denen Verbandsklagen gegen Diskriminierungen jetzt schon möglich sind, finden sie nur sehr selten statt. Die dazu auffindbaren Urteile[35] fallen vor allem durch ihre Exklusivität auf.

IV. Die neue Verbandsklagen-RL und ihre Umsetzung

Verbesserungen für die kollektive Rechtsdurchsetzung gegen verbrauchergefährdende KI könnte die europäische Verbandsklagen-Richtlinie[36] (Verbandsklagen-RL) versprechen. Diese muss in allen Mitgliedsstaaten der EU bis zum 25.12.2022 umgesetzt werden. Da die bisherigen kollektiven Rechtsschutzinstru-

---

[32] S. z. B. BAGE 147, 60; BAG NJW 2014, 1130.

[33] *Ponti/Tuchtfeld* ZRP 2018, 139; *Kocher* ZRP 2017, 112, 114; *Berghahn/Klapp/Tischbirek*, Evaluation des AGG, erstellt im Auftrag der Antidiskriminierungsstelle des Bundes, 2016, 161.

[34] *Berghahn/Klapp/Tischbirek* (Fn. 30), 148 ff. fordern die Streichung der Verschuldenserfordernisse in § 15 Abs. 1 S. 2, Abs. 3 AGG und § 21 Abs. 2 S. 2 AGG sowie der Beschränkung der Entschädigung auf drei Monatsgehälter in § 15 Abs. 2 Satz 2 AGG.

[35] So z. B. für eine Unterlassungsklage wegen Diskriminierung OLG Schleswig-Holstein VuR 2016, 190; für die Feststellungsklage eines Behindertenverbandes BVErwGE 125, 370.

[36] Richtlinie (EU) 2018/1828 über Verbandsklagen zum Schutz der Kollektivinteressen der Verbraucher und zur Aufhebung der Richtlinie 2009/22/EG.

mente noch nicht den Vorgaben der Verbandsklagen-RL entsprechen, wird auch Deutschland an dieser Stelle nachbessern müssen. Die wohl größte Innovation der Verbandsklagen-RL besteht darin, dass sie eine Verbandsklage vorsieht, die auf Abhilfe gerichtet ist (Art. 9 Abs. 1 der RL). Damit wird eine unmittelbar auf Leistung (z. B. in Form von Schadenersatz) gerichtete Kollektivklage ins deutsche Recht eingeführt. Der dringende Bedarf an dieser Klageform zeigt sich nicht zuletzt an den zahlreichen Legal Tech-Plattformen, die versuchen, in diese Lücke zu stoßen und die immer häufiger (vor allem von Verbrauchern) zur Rechtsdurchsetzung in Anspruch genommen werden.[37] In Bereichen, in denen bisher nur Unterlassungs- oder Feststellungsklagen möglich waren, wird das neue Möglichkeiten eröffnen.

1. Kein KI-spezifischer Anwendungsbereich

Dabei sollte jedoch klargestellt werden, dass die Verbandsklagen-RL nicht zu einer Erweiterung der materiellen Verbraucherrechte führt. Die Verbandsklagen-RL schafft keine eigenen verbraucherrechtlichen Ansprüche, sondern verweist über Art. 2 Abs. 1 auf ihren Anhang I, der zahlreiche unionsrechtliche Vorschriften enthält. Sie kann damit nur in Bereichen eingesetzt werden, in denen bereits materiell-rechtliche Schadenersatzansprüche der Verbraucher bestehen. Nichtsdestotrotz kommt den neuen Verbandsklagen ein weiter Anwendungsbereich zu, etwa für das Produkthaftungs- und -sicherheitsrecht, missbräuchliche Vertragsklauseln, Preisangaben oder Verstöße gegen das Wettbewerbsrecht. Auch hierunter fällt (mit Ausnahme der Regulierung des Geo-Blockings) keine „KI"-spezifische Regulierung.

Gegen diskriminierende Algorithmen kann beispielsweise das Wettbewerbsrecht einen Anknüpfungspunkt bieten. Wenn Marktteilnehmer unzulässigerweise diskriminierende Algorithmen einsetzen, kann dies einen Anspruch der Verbände nach § 8 Abs. 1 S. 1 UWG bzw. der Verbraucher nach § 9 Abs. 2 UWG n. F. begründen (s. dazu oben III.1.). Auf dieser Grundlage könnte eine Verbandsklage auf Abhilfe (Art. 9 Verbandsklagen-RL) oder Unterlassen (Art. 8 Verbandsklagen-RL) erfolgen.

Denkbar wäre freilich, zusätzliche Schadenersatzansprüche speziell gegen algorithmische Diskriminierung zu schaffen und diese in Anhang I der RL aufzunehmen – der Anhang ist durchaus erweiterungsfähig. Naheliegend wäre hier, die Vorschriften des KI-VO-E hinzuzufügen, sodass mittels der Verbandsklage eine Rechtsschutzmöglichkeit bei Verletzung der Zertifizierungsstandards gegeben wäre.

Primär kommt es allerdings darauf an, die Verbandsklagen-RL so umzusetzen, dass sie auch faktisch die funktionierende kollektive Rechtsdurchsetzung ermög-

---

[37] S. nur *Rühl* JZ 2020, 809, 812; *Fries* AcP 221 (2021), 109, 110 ff.

licht.[38] Konkret muss der Gesetzgeber ein effektives Verfahren schaffen, welches zügig und kostengünstig zu einem Leistungstitel führen kann. Zudem sollten der Kreis der klagebefugten Verbände und der sachliche Anwendungsbereich keinesfalls zu eng ausgestaltet sein.

2. Offenlegungspflicht für Beweismittel

Eine Besonderheit bringt noch Art. 18 Verbandsklagen-RL mit sich. Dieser ermöglicht, auch Beweismittel heranzuziehen, die der Verfügung des Beklagten oder eines Dritten unterliegen. Auf Antrag des klagenden Verbandes kann das Gericht eine Offenlegung dieser Beweismittel anordnen, wenn der Verband seinerseits alle unter zumutbarem Aufwand zugänglichen Beweismittel zur Stützung der Verbandsklage vorgelegt hat.

Gerade Darlegungs- und Beweisprobleme verhindern oft eine Rechtsdurchsetzung der von einem diskriminierenden Algorithmus Betroffenen, da diese regelmäßig keinen Einblick in die Funktionsweise eines Algorithmus haben. Die im deutschen Zivilprozessrecht de lege lata bestehenden Regeln zur Offenlegung von Beweismitteln helfen nur begrenzt weiter.[39] Zur Informationsgewinnung kann § 142 Abs. 1 ZPO nicht umfassend genutzt werden, da dazu die Urkunde schon hinreichend bezeichnet und in einem schlüssigen Vortrag in Bezug genommen worden sein muss.[40] Die §§ 421 ff., 428 ff., 432 ZPO enthalten Regelungen, die den Prozessgegner oder Dritte verpflichten, in ihrem Besitz befindliche beweiserhebliche Urkunden vorzulegen. Die Bestimmtheit des Antrags ist allerdings an hohe Anforderungen geknüpft, wie § 424 ZPO zeigt. Die Regelungen eignen sich daher nicht, um ein strukturelles Informationsdefizit auszugleichen.

Um Beweisschwierigkeiten bei Diskriminierungen entgegenzuwirken, sieht etwa § 22 AGG eine eigenständige Beweislastregel für Diskriminierungen vor. Für den Fall, dass eine Partei Indizien für eine Diskriminierung nachweist, trägt demnach die andere Partei die Beweislast dafür, dass keine Diskriminierung stattgefunden hat. Diese Regelung ist zweifellos nützlich,[41] gleichwohl bleibt der Rechtsschutz gegen diskriminierende Maßnahmen in der Praxis oft aus.[42]

Auch Art. 18 Verbandsklagen-RL setzt allerdings auf einer ähnlichen Stufe an. Die Offenlegungspflicht greift nur, wenn die klagende qualifizierte Einrichtung

---

[38] Erste Umsetzungsvorschläge s. *Gsell/Meller-Hannich*, Die Umsetzung der neuen EU-Verbandsklagenrichtlinie, 2021, abrufbar unter https://www.vzbv.de/sites/default/files/downloads/2021/02/03/21-02-04_vzbv_verbandsklagen-rl_gutachten_gsell_meller-hannich.pdf (Stand: 2.9.2022) sowie *Bruns*, Umsetzung der EU-Verbandsklagerichtlinie in deutsches Recht, 2022.

[39] Näher dazu *Hornkohl* GVRZ 2021, 17, Rn. 5 ff.; *Kern*, Urkundenvorlage bei Kartellschadensklagen, 2020, 24 ff.

[40] Vgl. Zöller/*Greger*, ZPO, 34. Auflage 2022, § 142 ZPO, Rn. 6 ff.

[41] Vgl. *Berghahn/Klapp/Tischbirek* (Fn. 30), 156 ff.

[42] *Braunroth* (Fn. 24), 456.

bereits hinreichend substantiiert vorgetragen hat. Erst dann sollen Beklagte oder Dritte zur Offenlegung weiterer Beweismittel verpflichtet sein – vorbehaltlich unionsrechtlicher und nationaler Vorschriften über Vertraulichkeit und Verhältnismäßigkeit. Da der Anspruch auch Dritte umfasst, kann z. B. auch vom Hersteller eines KI-Systems die Offenlegung der Funktionsweise des Systems verlangt werden. Das Problem in Fällen diskriminierender Algorithmen ist jedoch anders gelagert: Dadurch dass der Betroffene keine Möglichkeit hat, in den maßgeblichen Geschehensablauf Einblick zu nehmen, ist er schon meist nicht in der Lage, etwa die Diskriminierung oder die Intransparenz eines Algorithmus substantiiert vorzutragen. Gerade im Fall von intransparenten Algorithmen ist der fehlende Einblick in die Funktionsweise des Algorithmus das Kernproblem.[43]

Im Kartellrecht, wo Geschädigte auf vergleichbare Informationsasymmetrien treffen, besteht bereits eine Herausgabepflicht für Beweismittel nach §§ 33g, 89b ff. GWB.[44] Diese Regelungen könnten bei der Umsetzung von Art. 18 Verbandsklagen-RL als Vorbild dienen. Der Anspruch aus § 33g GWB ist nicht nur auf die Offenlegung im Prozess gerichtet, sondern auf Herausgabe der Beweismittel und Erteilung von Auskünften. § 33g Abs. 3 GWB schränkt den Anspruch ebenfalls im Hinblick auf die Verhältnismäßigkeit gegenüber dem Besitzer der Beweismittel ein.

Eine Alternative zeigt hingegen Art. 15 DSGVO auf, der der von einer Datenerhebung betroffenen Person ein Recht auf Auskunft darüber (u. a.), ob und zu welchem Zweck personenbezogene Daten verarbeitet werden, wenngleich der genaue Umfang der Auskunftspflicht strittig ist.[45] Eine weitere Alternative zum in Art. 18 Verbandsklagen-RL vorgesehen Weg wäre, umfangreiche Offenlegungspflichten für Beweismittel für solche Fälle einzuführen, in denen ein strukturelles Informations- und Ressourcengefälle zwischen den Parteien besteht und einer Partei ansonsten kein substantiierter Vortrag möglich ist.[46] Man könnte darüber hinaus an eine einsehbare Datenbank mit Traininsdaten und Zertifizierungen denken, so das Beweismittel auch für mehrere Prozesse einsehbar und nutzbar wären.

Dennoch ist auch die in Art. 18 Verbandsklagen-RL vorgesehene Möglichkeit, vom Beklagten oder von Dritten Beweismittel heraus zu verlangen, nicht zu unterschätzen. Sie könnte im Vorfeld mit dem Ansatz der sekundären Darlegungslast kombiniert werden, der gerade in den „Diesel-Klagen" von der Rechtsprechung bereits vielfältig angewendet wurde, um Darlegungs- und Beweisschwierigkeiten der Kläger entgegenzuwirken.[47] Voraussetzung dafür ist, dass der eigentlich pri-

---

[43] *Hornkohl* (Fn. 36), Rn. 30.

[44] Der deutsche Gesetzgeber setzte damit die Vorgaben nach Art. 5 der Kartellschadenersatzrichtlinie um; zur Bewertung der Umsetzung *Kern* (Fn. 36), 133 ff.

[45] Dazu BeckOK-Datenschutzrecht/*Schmidt-Wudy*, 37. Edition Stand 1.8.2021, Art. 15 DSGVO, Rn. 52 ff.

[46] S. dazu *Hornkohl* (Fn. 36), Rn. 30 ff.; vgl. auch *Kern* (Fn. 36), Kapitel 4 speziell für das Kartellrecht.

[47] Z. B. bei BGH NJW 2020, 2804, 2805 f.; NJW 2020, 1962, 1966 f.

mär darlegungspflichtige Kläger außerhalb des von ihm dazulegenden Geschehensablaufs steht und keine nähere Kenntnis der maßgeblichen Tatsache hat und sich diese auch nicht verschaffen kann, während der Beklagte Kenntnis davon hat und ihm die näheren Angaben zumutbar sind.[48] Dies sollte in Fällen diskriminierender Algorithmen regelmäßig der Fall sein. In diesem Fall muss der Beklagte diese Angaben selbst vornehmen, andernfalls die Behauptungen des Klägers genügen.

V. Fazit

Während einerseits Fälle von diskriminierenden Algorithmen wegen deren regelmäßig massenhaften Einsatzes geradezu exemplarisch geeignet für Verbandsklagen sind, gibt es andererseits bisher keine „KI“-spezifischen Verbandsklagen. Die bestehenden kollektiven Rechtsschutzmöglichkeiten erfassen die besondere Gefährdung durch diskriminierende Algorithmen nicht hinreichend. Hier sollten entsprechende Klagemöglichkeiten geschaffen werden, etwa indem die Vorschriften des KI-VO-E in Anhang I der Verbandsklagen-RL aufgenommen werden. Für die Effektivität dieses Weges wird es freilich entscheidend auf die Art und Weise der Umsetzung der Verbandsklagen-RL ins nationale Recht ankommen.

[48] BGH NJW 2014, 2360, 2361; NJW 1990, 1351, 1352; NJW 1987, 1201; MüKo-ZPO/*Fritsche*, 6. Auflage 2020, § 138 ZPO, Rn. 24.

Algorithmische Entscheidungssysteme im Nichtdiskriminierungsrecht

Dogmatische Herausforderungen und konzeptionelle Perspektiven

Jan-Laurin Müller[1]

I. Einleitung

Angetrieben von technischem Fortschritt in den Bereichen *Machine Learning* und *Big Data* trifft sog. „Künstliche Intelligenz" vermehrt nichttriviale Entscheidungen mit erheblichen Auswirkungen auf weite Teile des menschlichen Lebens. Die Verheißungen solcher *Algorithmic Decision Making-Systems*[2] (kurz: ADM-Systeme) sind groß. Sie sollen Effizienz und Fairness von Entscheidungsfindungsprozessen gleichermaßen fördern:[3] Ersteres, weil sie – anders als menschliche Entscheidungsträger*innen – akkurate Datenmodelle verwenden und alle für die Entscheidung relevanten Faktoren einbeziehen.[4] Letzteres, indem sie *menschlichen bias* unterbinden,[5] und durch verstärkte Differenzierung eine unterschiedliche Behandlung ungleicher Sachverhalte gewährleisten.[6] Auf diese Weise werde eine Grundlage für „rationale Differenzierungen"[7] geschaffen und damit das Ende

---

[1] Der Beitrag basiert im Wesentlichen auf ersten Gedanken meines Promotionsvorhabens (work in progress) und dem von mir verantworteten Teil des Vortrags, den ich am 12.7.2021 gemeinsam mit *Michael Grünberger* auf den Verbraucherrechtstagen gehalten habe. Ich danke *ihm* sowie *Ruth Janal* für wertvolle Hinweise. Sämtliche online verfügbaren Inhalte wurden zuletzt aufgerufen am 10.11.2021.

[2] Für die Zwecke des vorliegenden Beitrags werden darunter solche Entscheidungssysteme verstanden, die auf irgendeiner Art des maschinellen Lernens aufbauen, das heißt im Wege eines Lernprozesses aus Daten (mehr oder weniger selbstständig) ein Entscheidungsmodell destillieren. Zur grundlegenden Funktionsweise solcher Systeme siehe *Russel/Norvig*, Artificial Intelligence: A Modern Approach, 4. Aufl. 2021, 1–31 und 651–822; *Shalev-Shwartz/Ben-David*, Understanding Machine Learning: From Theory to Algorithms, 2014; (mit Fokus auf verschiedenen Techniken des *Machine Learning*); *Calders/Custers*, in: Custers et al. (Hrsg.), Discrimination and Privacy in the Information Society, 2013, 27–42 (mit Fokus auf Techniken des *Data Mining*).

[3] Paradigmatisch für die (präsumtive) Überlegenheit statistischer Entscheidungen *Dawes/Faust/Meehl*, Science (New Series) 1989, 1668.

[4] Dazu *Zarsky*, I/S: A Journal of Law and Policy for the Information Society 2017, 11 (13).

[5] Dazu *Zarsky*, Science, Technology, & Human Values 2016, 118 (122); *Zarsky*, (Fn. 4), S. 13. Zu menschlichen *biases* siehe *Kahneman/Sibony/Sunstein*, Noise: A Flaw in Human Judgement, 2021.

[6] Berechtigte Kritik dazu bei *Zarsky*, Washington Law Review 2014, 1375 (1382).

[7] Zum Konzept *Gandy Jr.*, Coming to Terms with Chance. Engaging Rational Discrimination and Cumulative Disadvantage, 2016, S. 55–76.

jeglicher Diskriminierung eingeläutet.[8] Die Wirklichkeit zeichnet, wie so häufig, ein komplexeres Bild: Wo Licht ist, da fällt bekanntlich auch Schatten. Unumstritten ist mittlerweile wohl, dass ADM-Systeme ein erhebliches Diskriminierungspotential haben, was nicht nur die mediale Berichterstattung,[9] sondern auch die empirische Forschung[10] nahelegt.

Paradigmatisch für dieses Diskriminierungspotential stehen zwei Beispiele der jüngeren Vergangenheit:[11] Im ersten Fall geht es um Diskriminierung bei der Kreditvergabe durch den Online-Finanzdienstleister *Svea Ekonomi*.[12] Das schwedische Unternehmen nutzte ein ADM-System zur Entscheidung über Anträge auf Kreditvergabe. Den entscheidungserheblichen Kreditscore ermittelte der Algorithmus dabei unter anderem anhand der Kriterien Geschlecht, Alter, Muttersprache und Wohnort. Der Antrag eines finnisch sprechenden, männlichen Antragsstellers wurde aufgrund seines zu niedrigen Scores abgelehnt. Das finnische nationale Nichtdiskriminierungs- und Gleichheitstribunal verurteilte den Finanzdienstleister deshalb im Jahr 2018 mit der Begründung, Svea Ekonomi habe die negative Entscheidung an unzulässige, diskriminierende Kriterien geknüpft, anstatt eine umfassende Einzelfallprüfung durchzuführen. Der zweite Fall betrifft einen von *Amazon* entwickelten Recruitment-Algorithmus.[13] Amazon testete ein ADM-System, um geeignete Bewerber*innen für Stellen im IT-Bereich anhand von online verfügbaren Lebensläufen ausfindig zu machen. Schon bald zeigte sich aber, dass der Algorithmus potenzielle Bewerber*innen nicht geschlechtsneutral beurteilte, sondern Frauen systematisch eine signifikant schlechtere Bewertung zuwies. Grund dafür waren unausgewogene Trainingsdaten: Der Algorithmus wurde anhand von personenbezogenen Daten der bestehenden Belegschaft trainiert, die überwiegend aus Personen männlichen Geschlechts bestand. Auf diese

---

[8] Zu einigen dieser Verheißungen statistischer Entscheidungsfindungsprozesse siehe *Dawes/Faust/Meehl* (Fn. 1). Kritische Beurteilung bei *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, S. 20–23 et passim.

[9] Exemplarisch *Angwin* et al., ProPublica vom 23.5.2016, abrufbar unter https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; *Dastin*, Reuters Online vom 11.10.2018, abrufbar unter https://www.reuters.com/article/us- amazon-com-jobs-automation-insight-idUSKCN1MK08G; *Rähm*, Deutschlandfunk vom 24.3.2019, abrufbar unter https://www.deutschlandfunk.de/algorithmen-im-arbeitsamt-wenn-kuenstliche-intelligenz.724.de.html?dram:article_id=444465.

[10] Exemplarisch *Kim* et al., arXiv:2005.05921, Edition 28.5.2020 (Hate Speech); *Suriyakumar* et al., Conference on Fairness, Accountability, and Transparency 2021, 723 (Health Care). Umfangreiche Nachweise auch bei *O'Neil*, Weapons of Math Destruction, 2016 und *Orwat* (Fn. 8).

[11] Wie in vielen anderen bekannten Beispielsfällen ist für Beobachtende unklar, ob und wenn ja welche Techniken maschinellen Lernens konkret eingesetzt werden. Für die Zwecke des vorliegenden Beitrags unterstelle ich daher, die Systeme würden auf solchen Verfahren beruhen. Abweichende Ergebnisse ergeben sich durch diese Prämisse nicht.

[12] Beispiel nach *National Non-Discrimination and Equality Tribunal of Finland*, Register Number 216/2017, abrufbar unter https://www.yvtltk.fi/material/attachments/ytaltk/tapausselosteet/45LI2c6dD/YVTltk-tapausseloste-_21.3.2018-luotto-moniperusteinen_syrjinta-S-en_2.pdf.

[13] Beispiel nach *Dastin* (Fn. 9).

Weise lernte er, die Variable „weiblich" bzw. damit korrelierende Stellvertretermerkmale negativ zu bewerten. Beide Beispiele verdeutlichen das erhebliche Diskriminierungspotenzial, das von ADM-Systemen ausgeht, wenn sie auf Verfahren maschinellen Lernens basieren.

Dieses Diskriminierungspotenzial wurde zwischenzeitlich auch auf legislativer Ebene erkannt und ein entsprechender Handlungsbedarf identifiziert: Die Europäische Kommission hat zunächst ein Weißbuch[14] zur Regulierung „Künstlicher Intelligenz" vorgelegt, zu dem sowohl die 24. Bundesregierung[15] wie auch die Koalitionsfraktionen des 19. Deutschen Bundestages[16] Stellung bezogen haben. Den vorläufigen Schlusspunkt dieser regulatorischen Debatte bildet der von der Europäischen Kommission vorgelegte Vorschlag für einen „AI-Act",[17] der ebenso wie das Weißbuch die Diskriminierungsfreiheit von ADM-Systemen als eine zentrale Herausforderung der KI-Regulierung einstuft.[18] Auffällig ist dabei, dass der Vorschlag zwar das besondere Diskriminierungspotenzial von ADM-Systemen (an-) erkennt, aber dennoch gerade keine spezifischen Vorgaben für das europäische Nichtdiskriminierungsrecht macht. Er verfolgt stattdessen einen risikobasierten Ansatz. Das Nichtdiskriminierungsrecht fehlt dagegen bislang als ausdrückliches Instrument im KI-Regulierungskonzept der Europäischen Union. Dies wirft die zentrale Frage nach der Zukunft des Rechtsgebiets in einer zunehmend digitalen Gesellschaft auf: Welche Rolle kann und sollte das deutsche und europäische Nichtdiskriminierungsrecht[19] bei der Regulierung diskriminierender ADM-Systeme übernehmen? Dieser Beitrag entwickelt die These, dass die spezifische Form der Diskriminierung durch ADM-Systeme die tragenden Strukturen und Konzepte des Nichtdiskriminierungsrechts nachhaltig herausfordert. Die entstehenden Diskriminierungsrisiken können zwar teilweise durch eine Anpassung der bestehenden dogmatischen Instrumente adressiert werden. Gleichzeitig werden

---

[14] Europäische Kommission, Weißbuch Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen, KOM(2020)65 endg.

[15] Stellungnahme der Bundesregierung der Bundesrepublik Deutschland zum Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen, abrufbar unter https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2020/stellungnahme-breg-weissbuch-ki.pdf.

[16] Antrag der Fraktionen CDU/CSU und SPD. Zukunftstechnologie Künstliche Intelligenz als Erfolgsfaktor für ein starkes und innovatives Europa – Eine Stellungnahme zum Weißbuch „Zur Künstlichen Intelligenz" der EU-Kommission, BT-Drs. 19/22181.

[17] 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, KOM(2021)206 endg.

[18] Zum ‚Weißbuch' siehe dort v.a. die Seiten 1 und 11. Zum ‚AI-Act' siehe zentral ErwG (15) sowie die ErwG (17), (28), (35–37) und (39).

[19] Zum europäischen und deutschen Nichtdiskriminierungsrecht werden im Folgenden all diejenigen unionalen und deutschen Rechtsvorschriften gezählt, die Diskriminierungsverbote im Horizontalverhältnis etablieren. Dazu zählen insbesondere Art. 157 AEUV, die Richtlinien 2000/43/EG, 2000/78/EG, 2006/54/EG, 2004/113/EG, die Richtlinienvorschläge KOM(2008), 426 und KOM(2021), 93, sowie das deutsche AGG.

aber auch ganz neue konzeptionelle wie dogmatische Perspektiven notwendig, um der besonderen Form und den spezifischen Risiken algorithmischer Diskriminierung gerecht zu werden. Die Begründung dieser These erfolgt in drei Schritten: Zunächst (II.) erläutere ich, welche Eigenschaften die spezifische Form der Diskriminierung durch ADM-Systeme charakterisieren und welche besonderen Diskriminierungsrisiken damit verbunden sind. In einem zweiten Schritt (III.) zeichne ich kurz nach, wie das Nichtdiskriminierungsrecht diesen Risiken bereits heute begegnen kann und welche dogmatischen Hürden dabei auftreten (können). Zuletzt (IV.) wage ich einen Blick in die Zukunft und gehe der Frage nach, ob die tragenden Konzepte und Strukturen des gegenwärtigen Nichtdiskriminierungsrechts noch geeignet sind, um die spezifische Form ADM-basierter Diskriminierung rechtlich adäquat abzubilden.

II. Die spezifische Form der Diskriminierung durch ADM-Systeme

Was charakterisiert also die spezifische Form, in der Diskriminierung durch ADM-Systeme auftritt? Was sind ihre Ursachen (II.1.) und was die damit verbundenen besonderen (Diskriminierungs-)Risiken (II.2.)?

1. Ursachen der Diskriminierung durch ADM-Systeme

Aus sozio-technischer Perspektive kann Diskriminierung durch ADM-Systeme auf einer Vielzahl von Gründen beruhen. Zur Systematisierung möchte ich eine vierteilige Matrix vorschlagen, die zwischen *biased training data, unequal base rates, bias in modeling* und *bias in usage* unterscheidet.[20]

Zur ersten Fallgruppe der *biased training data* gehören all diejenigen Fälle, in denen das algorithmische Datenmodell an irgendeiner Art von „Fehler" oder „Mangel" leidet. Die Gründe dafür sind vielfältig:[21] (1.) Bereits die Operationalisierung der Zielvorgabe des ADM-Systems durch eine mathematisch präzise Definition von Zielvariablen ist fehleranfällig für *biases*[22] (*biased definition of target variab-*

[20] Alternative Systematisierungsansätze finden sich bei *Kleinberg* et al., Journal of Legal Analysis 2018, 113 (138–146): Fokus auf Fehlern im Umgang mit Trainingsdaten; *Hacker*, Common Market Law Review 2018, 1143 (1146–1150): zweiwertige, technische Unterscheidung; *Barocas*/*Selbst*, California Law Review 2016, 671 (677–693): ausdifferenzierte fünfwertige, technische Unterscheidung; *Ferrer* et al., IEEE Technology and Society Magazine 2021, 72 (72–73): dreiwertige Unterscheidung von *bias in modeling*, *bias in training* und *bias in usage*; *Xenidis*/*Senden*, in: Bernitz et al. (Hrsg.), General Principles of EU Law and the EU Digital Order, 2020, 151 (156–160): zweiwertige, sozialwissenschaftlich-normative Systematisierung.

[21] Grundlegend *Barocas*/*Selbst* (Fn. 20), S. 677–693.

[22] Der Begriff des *bias* wird in diesem Abschnitt grundsätzlich nicht in einem normativen Sinn verstanden. Es wird seine statistische Bedeutung zugrunde gelegt, also die Frage gestellt, ob ein Erwartungswert oder Durchschnittswert vom wahren Wert der zu beurteilenden Größe abweicht.

les).[23] Beim Amazon-Recruitment Algorithmus macht es beispielsweise einen entscheidenden Unterschied, mit Hilfe welcher Zielvariable(n) die Eigenschaft gemessen wird, ob eine Person eine „gute" Arbeitnehmer*in ist: Produktivität, Betriebszugehörigkeit oder Lohnkosten etc.[24] (2.) Auch die Auswahl und Zusammenstellung des Trainingsdatensatzes kann zu „Fehlern" des algorithmischen Datenmodells führen, und das in zweifacher Hinsicht: Zum einen, wenn das ADM-System seiner Funktionsweise nach darauf abzielt, vergangene menschliche Entscheidungen zu rekonstruieren, die ihrerseits diskriminierend waren (*historical bias*).[25] Paradigmatisch für diese Fallgruppe steht wiederum der Amazon-Recruitment Algorithmus, der mit Datensätzen gefüttert wurde, die auf menschlichen Einstellungsentscheidungen beruhten, von denen primär Männer profitiert hatten. Zum anderen können „mangelhafte" Trainingsdatensätze darauf beruhen, dass nicht ausreichend Daten über bestimmte gesellschaftliche Gruppen verfügbar sind oder die Menge, Qualität oder Aktualität der Daten zwischen den Gruppen variiert (*sampling bias*).[26] (3.) Drittens ist es möglich, dass die ins Datenmodell einbezogenen Variablen unterschiedlich adäquate Aussagen für verschiedene soziale Gruppen zulassen (*feature selection bias*).[27] Die Frage, welche Merkmale eines Datensatzes in das Datenmodell einbezogen werden, erlangt damit entscheidende Bedeutung. Im Fall der ADM-basierten Kreditvergabe kann das zum Beispiel dann zutreffen, wenn das Kriterium des Wohnortes für bestimmte Gruppen (Männer oder Frauen) unterschiedlich zuverlässig Informationen über die Kreditwürdigkeit bereitstellt.

Aber selbst wenn das algorithmische Datenmodell frei von „Mängeln" und „Fehlern" sein sollte, besteht die Möglichkeit, dass es eine faktisch zutreffende, aber normativ problematische Abbildung bestehender, gesellschaftlicher Ungleichheiten zwischen verschiedenen Individuen und Gruppen beinhaltet (*unequal base rates*[28] oder *unequal ground truth*[29]). Die Ungleichbehandlung ist hier gerade nicht das Produkt einer unzutreffenden oder verzerrenden Abbildung der Realität wie im Fall der *biased training data*. Sie beruht vielmehr darauf, dass das ADM-System eine in verschiedenen gesellschaftlichen Teilsystemen bestehende Ungleichheit aufgreift und anhand dieser seine eigene, technische Normativität bestimmt. Damit verstärkt das System das normative Problem der Ungleichheit, weil es diese in die Zukunft hinein fortschreibt.

---

[23] Dazu *Barocas/Selbst* (Fn. 20), S. 677–680.

[24] Zu diesem Beispiel siehe *Barocas/Selbst* (Fn. 20), S. 679.

[25] Dazu *Calders/Žliobaitė*, in: Custers et al. (Hrsg.), Discrimination and Privacy in the Information Society, 2013, 43 (50). Ähnlich aber mit leicht anderer Akzentuierung *Barocas/Selbst* (Fn. 20), S. 681–684.

[26] Dazu *Calders/Žliobaitė* (Fn. 25), S. 50–51. Ähnlich aber mit leicht anderer Akzentuierung *Barocas/Selbst* (Fn. 20), S. 684–688.

[27] Dazu *Barocas/Selbst* (Fn. 20), S. 688–690. Ähnlich *Calders/Žliobaitė* (Fn. 25), S. 52–53.

[28] Begriff nach *Barocas/Hardt/Narayanan*, Fairness and Machine Learning. Limitations and Opportunities, Edition 16.6.2021, 23.

[29] Begriff nach *Hacker* (Fn. 20), S. 1148–1150.

Zur dritten Fallgruppe des *bias in modeling*[30] gehören sämtliche Eingriffe von Entwickler*innen in das algorithmische Datenmodell. Die Motive dafür sind vielfältig: Zum einen können Entwickler*innen versuchen, statistische Ungleichheiten, die auf *biased training data* oder *unequal base rates* beruhen, zu beseitigen (*algorithmic processing bias*[31]). Die Eingriffe können aber auch auf die Verwirklichung sonstiger extern-normativer Vorgaben durch Recht oder Moral abzielen (*algorithmic focus bias*[32]). Diskriminierende Ergebnisse entstehen bei dieser Fallgruppe häufig dadurch, dass in der Regel ein *trade-off* zwischen verschiedenen (rechtlichen, ethischen und statistischen) *fairness*-Begriffen besteht und die Beseitigung des einen *bias* regelmäßig mit der Produktion anderer *biases* einhergeht.[33]

Zuletzt kann Diskriminierung auch erst bei der Anwendung des ADM-Systems entstehen (*bias in usage*[34]). Dies wird häufig der Fall sein, wenn ein algorithmisches Datenmodell für einen bestimmten sozialen Kontext entwickelt wurde und anschließend ohne ausreichende Anpassungen auf einen neuen sozialen Kontext übertragen wird (*decontextualization bias* oder *transfer context bias*[35]). Man denke nur an die Verwendung von *credit-scores* in verschiedenen sozialen Kontexten, um die Vertrauenswürdigkeit einer Person zu messen.[36] Ferner kann eine Diskriminierung im Anwendungsprozess auch dadurch entstehen, dass (menschliche) Anwender*innen oder weitere autonome Systeme, die mit den Ergebnissen des ADM-Systems gefüttert werden, das Ergebnis des algorithmischen Entscheidungsprozesses miss- bzw. fehlinterpretieren (*interpretation bias*[37]).

Allen vier Fallgruppen gemeinsam ist, dass jeweils die Frage der „richtigen" Konstruktion bzw. des „richtigen" Umgangs mit dem Datenmodell des ADM-Systems im Zentrum steht. Charakteristisch für die spezifische Form ADM-basierter Diskriminierung ist demnach ihre Datenbezogenheit. Sie determiniert in soziotechnischer Hinsicht die vier Koordinaten der Ursachen algorithmischer Diskriminierung. Diese Systematisierung lässt sich im Recht für zwei wichtige dogmatische Fragestellungen fruchtbar machen: (1.) Welche Anforderungen sind an die Rechtfertigung einer Ungleichbehandlung zu stellen?[38] (2.) Wie sind die Verant-

---

[30] Begriff nach *Ferrer* et al. (Fn. 20), S. 73.

[31] Begriff nach *Danks*/*London*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence 2017, 4691 (4693).

[32] Begriff nach *Danks*/*London*, (Fn. 31), S. 4693.

[33] Grundlegend zu dieser Erkenntnis für statistische *bias*-Begriffe *Chouldechova*, arXiv:1610.07524v1, Edition 24.10.2016. Vgl. exemplarisch zum *bias-variance trade-off*: *James* et al., An Introduction to Statistical Learning, 2013, 33–36; *Geman*/*Bienenstock*/*Doursat*, Neural computation 4 (1992), 1–58.

[34] Begriff nach *Ferrer* et al. (Fn. 20), S. 73.

[35] Dazu *Danks*/*London* (Fn. 31), S. 4694.

[36] Beispiele dazu bei *O'Neil* (Fn. 10), S. 7 und 147–150.

[37] Dazu *Danks*/*London* (Fn. 31), S. 4694.

[38] Zu diesem Aspekt auf Grundlage seiner zweiwertigen Unterscheidung auch *Hacker* (Fn. 20), S. 1160–1167.

wortlichkeiten zuzuweisen, insbesondere: Wer haftet für die Diskriminierung? Das möchte ich kurz am Beispiel der Rechtfertigungsproblematik skizzieren, weil sich dort zeigt, dass sich die vier Fallgruppen sowohl hinsichtlich des Bezugspunktes der Rechtfertigungsprüfung als auch hinsichtlich der dogmatischen Verankerung innerhalb der Rechtfertigungsebene unterscheiden: (1.) Im Fall von *biased training data* wird es zumeist um den Vorwurf der Verwendung ungeeigneter Daten oder der ungeeigneten Strukturierung von Daten gehen. Dogmatisch gewandet betrifft das die Geeignetheit bzw. Erforderlichkeit der ungleichen Behandlung. (2.) Bei Fällen der *unequal base rates* steht der Vorwurf einer normativ problematischen Abbildung diskriminierender Realitäten im Raum. Das betrifft rechtsdogmatisch primär die Angemessenheitsprüfung. (3.) *Biases in modeling* werfen die spannende Frage auf, ob die Verwirklichung eigener ethisch fundierter Fairness-Vorstellungen durch private Akteure unter den dogmatischen (Rechtfertigungs-[39])Tatbestand der Positivmaßnahmen gefasst werden können und ob diese Eingriffe geeignet, erforderlich und angemessen sind. (4.) In den Fällen des *bias in usage*, vor allem beim *decontextualization bias* geht es um die Leistungsfähigkeit des ADM-Modells. Rechtsdogmatisch handelt sich also um die Frage der Geeignetheit und Erforderlichkeit innerhalb der Rechtfertigungsprüfung.

2. Diskriminierungsrisiken beim Einsatz von ADM-Systemen

Welche sind nun die sozialen Risiken, die diese spezifische Form datenbezogener Diskriminierung durch ADM-Systeme mit sich bringt? Was sind ihre sozio-technischen Hintergründe, ihre sozio-ökonomischen Ursprünge, und worin besteht ihre normative Bedenklichkeit? Die Vielschichtigkeit einer Antwort auf diese Frage(n) steht der Komplexität der zugrundeliegenden Systeme in nichts nach: Die Liste reicht von Risiken für Schäden an individuellen Rechtsgütern wie Leib, Leben und Vermögen,[40] über gesellschaftliche Risiken für die Funktionsfähigkeit des demokratischen Verfassungsstaates[41] bis hin zu spezifischen Diskriminierungsrisiken der Verfestigung alter und der Schaffung neuer Ungleichheiten.[42] Für die Frage nach der Relevanz des Nichtdiskriminierungsrechts bei der Regulierung von ADM-Systemen sind vor allem diese letzten, *spezifischen Diskriminierungsrisiken* von Bedeutung. Den vielstimmigen Diskurs stark vereinfachend und

---

[39] Zum Streitstand um die dogmatische Einordnung von Positivmaßnahmen siehe BeckOGK-AGG/*Baumgärtner*, Edition 15.9.2021, § 5 Rn. 10.

[40] Instruktiv zur Problemlage *Awad* et al., Nature 563 (2018), 59.

[41] Exemplarisch Unger/von Ungern-Sternberg (Hrsg.), Demokratie und künstliche Intelligenz, 2019.

[42] Paradigmatisch dafür aus der reichhaltigen Literatur *O'Neil* (Fn. 10); *Noble*, Algorithms of Oppression. How Search Engines Reinforce Racism, 2018; *Eubanks*, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, 2018; *Perez*, Invisible Women: Exposing Data Bias in a World Designed for Men, 2019; *Benjamin*, Race After Technology: Abolitionist Tools for the New Jim Code, 2019; *D'Ignazio/Klein*, Data Feminism, 2020; *Chun*, Discriminating Data, 2021.

systematisierend, lassen sich vier übergreifende solcher Risiken[43] identifizieren:[44] Perpetuierungsrisiken, Generalisierungsrisiken, Korrelationsrisiken und Transparenzrisiken.

a) Perpetuierungsrisiko

Mit *Perpetuierungsrisiko*[45] beschreibe ich den Effekt, dass ADM-Systeme bestimmte soziale (insbesondere ökonomische) Gesellschaftsstrukturen aufgreifen, ihre eigene technisch-normative Handlungslogik nach diesen Strukturen ausrichten und diese damit verfestigen. *Sozio-technischer Hintergrund* dieses Effekts ist, dass ADM-Systeme die entsprechenden Gesellschaftsstrukturen in den ihnen zugrunde liegenden Datenmodellen abbilden und dadurch als Beurteilungsgrundlage für gegenwärtige und künftige Entscheidungen festlegen. Die abgebildeten Strukturen erhalten auf diese Weise eine technisch-normative Verbindlichkeit.[46] Diese Verbindlichkeit weist sowohl eine vergangenheitsbezogene wie auch eine zukunftsbezogene Dimension auf: Erstere wird vor allem dann relevant, wenn ADM-Systeme ihre Modelle anhand von vergangenheitsbezogenen Daten konstruieren und gegenwärtige Entscheidungen danach ausrichten. Dies wird insbesondere beim Amazon-Recruitment-Algorithmus deutlich, dessen Datenmodell vergangene, menschliche Diskriminierungen abgebildet hat. Die zukunftsbezogene Dimension der technisch-normativen Verbindlichkeit zeigt sich, wenn die Entscheidungen des ADM-Systems – im Sinne eines performativen Aktes[47] – selbst Einfluss auf die Trainingsdatenbasis zukünftiger Modelle oder auf die zukünftigen Handlungsoptionen der betroffenen Personen nehmen.[48] Seinen

---

[43] Von *Risiken* anstatt von bloßen *Effekten* möchte ich aus zwei Gründen sprechen: Erstens, weil alle beschriebenen Effekte auch in normativer Hinsicht problematisch sind. Zweitens, weil der Einsatz von ADM-Systemen viele der Effekte (im Gegensatz zu menschlichen Entscheidungsprozessen) kontrollierbar(er) macht. Zu diesem letzten Aspekt vgl. den Risikobegriff nach *Luhmann*, Soziologie des Risikos, 2003.

[44] Alternative Systematisierung bei *Orwat* (Fn. 8), S. 85–96, der insgesamt sechs verschiedene Risiken unterscheidet und dabei die normativ-rechtliche Bewertung teilweise von der Beschreibung der faktischen Problemlage trennt.

[45] Instruktiv und mit Beispielen dazu *O'Neil* (Fn. 10), S. 7, 27, 42, 53–54, 80–81, 87, 112, 129, 155, 167: „feedback loops"; *Noble* (Fn. 42). Allgemein dazu bei statistischen Differenzierungsprozessen *Britz*, Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung, 2008, 125–130: „Akkumulationseffekte"; siehe auch *Gandy Jr.*, Ethics and Information Technology 2010, 29 (37–39). Aus rechtlicher Perspektive im ADM-Kontext *Citron/Pasquale*, Washington Law Review 2014, 1 (32–33): „negative spirals"; *Zarsky*, (Fn. 6), S. 1405–1408: „negative spiral"; *Orwat* (Fn. 8), S. 89–90: „Akkumulations- und Verstärkungseffekte"; *Xenidis*, Maastricht Journal of European and Comparative Law 2020, 736 (751–753); *Xenidis/Senden* (Fn. 20) S. 158–160; *Gerards/Borgesius*, Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence, Edition 6.7.2021, 8–9, abrufbar unter https://ssrn.com/abstract=3723873.

[46] Ähnlich *Hildebrandt*, in: Pelillo/Scantamburlo (Hrsg.), Machines We Trust. Perspectives on Dependable AI, 2021, 56: „*code-driven normativies*".

sozio-ökonomischen Ursprung hat der Perpetuierungseffekt darin, dass ADM-Systeme es ermöglichen, „verborgene Gesellschaftsstrukturen" sichtbar und operationalisierbar zu machen, indem sie in ihren Datenmodellen Korrelationsbeziehungen zwischen bestimmten Kategorien sozialer Kommunikation aufdecken. Sie erfüllen damit genau diejenige Funktion, die der Soziologe *Armin Nassehi* aus systemtheoretischer Sicht für den Siegeszug der Digitaltechnik verantwortlich macht:[49] Durch das Sichtbarmachen „verborgener Gesellschaftsstrukturen" ermöglichen es ADM-Systeme ihren Anwender*innen, bislang unentdeckte Effizienzvorteile dieser Strukturen zu heben.

Dieser Effekt ist in normativer Hinsicht bedenklich: Aus *normativ-ethischer Sicht* wird dies besonders deutlich, wenn man konsequentialistische Ansätze zur Legitimierung des Nichtdiskriminierungsrechts in den Blick nimmt. Gemeinsame Grundlage dieser Ansätze ist der Gedanke, dass sich der ethische Unwert von Diskriminierung daraus ergibt, dass gesellschaftliche Strukturen der Hierarchisierung und Exkludierung gefördert und aufrechterhalten werden. Diese Hierarchisierung und Exklusion kann sich ergeben aus einer Beeinträchtigung symbolischer Repräsentation (Förderung symbolischer Ungleichheit)[50] oder einer Beeinträchtigung des Zugangs zu materiellen wie immateriellen Gütern (Förderung distributiver Ungleichheit)[51].[52] Für beide Ansatzpunkte liefert der Einsatz von ADM-Systemen ausreichend Fallmaterial: Hinsichtlich der Förderung distributiver Ungleichheit denke man nur an die Eingangsfälle einer ADM-basierten Kreditvergabe oder eines ADM-basierten Bewerbungsverfahrens. Die Fälle diskriminierender Suchmaschinen[53] und Textverarbeitungssoftware[54] stehen stellvertretend für die Förderung symbolischer Ungleichheit. Aus *normativ-rechtlicher Perspektive* sind Perpetuierungseffekte sowohl unter freiheitsrechtlichen wie auch unter gleichheitsrechtlichen Gesichtspunkten problematisch. Aus freiheitsrechtlicher Perspektive bedeuten sie eine Beeinträchtigung der grundrechtlich durch Art. 2 Abs. 1 GG i. V. m. Art. 1 Abs. 1 GG geschützten freien Entfaltung der Persönlichkeit.[55] Auch vor dem Hintergrund eines gleichheitsrechtlichen Paradigmas erscheinen sie problematisch, weil es nach (zwar nicht unbestrittener[56]) aber

---

[47] *Xenidis* (Fn. 45), S. 751: „performativity of predictive analytics"; *Xenidis*/*Senden* (Fn. 20), S. 160: „The effect is performative.".

[48] Dazu mit Beispielen *Barocas*/*Hardt*/*Narayanan* (Fn. 28), S. 25–26.

[49] *Nassehi*, Muster. Theorie der digitalen Gesellschaft, 2019.

[50] Grundlegend dazu *Fraser*, New Left Review 1995, 68.

[51] Exemplarisch dazu *Dworkin*, Philosophy & Public Affairs 1981, 185, 283.

[52] Siehe zum Ganzen auch das vierdimensionale Konzept materialer Gleichheit bei *Fredman*, Discrimination Law, 2. Auflage 2011, 25–33.

[53] Dazu *Noble* (Fn. 42).

[54] Dazu erstmalig aus technisch-sozialwissenschaftlicher Sicht *Caliskan*/*Bryson*/*Narayanan*, Science 356 (2017), 183.

[55] Allgemein dazu bei statistischen Differenzierungsprozessen *Britz* (Fn. 45), S. 179–209.

[56] Paradigmatisch *Lobinger*, AcP 216 (2016), 28 (80–101).

gerade auch rechtsvergleichend überwiegender Auffassung[57] Aufgabe und Ziel des Nichtdiskriminierungsrechts ist, bestehende Strukturen sozialer Hierarchisierung zumindest in Frage zu stellen, wenn nicht gar zu überwinden.

b) Generalisierungsrisiko

Der Begriff des *Generalisierungsrisikos*[58] adressiert die Kontextualisierungs- und Individualisierungsdefizite algorithmischer Entscheidungsprozesse. Das ist ein klassisches Problem aller statistischen Differenzierungsprozesse, das sich aus *sozio-technischer Sicht* im Fall ADM-basierter Diskriminierung zuspitzt: Weil einzelne Individuen als eine Kumulation von Zugehörigkeiten zu unterschiedlichen (Kategorien von) sozialen Gruppen rekonstruiert werden, kommt es zu statistisch validen, aber im konkreten Einzelfall potentiell unzutreffenden Aussagen über einzelne Personen. Statistische Annahmen, die die Mehrzahl der Angehörigen einer bestimmten Gruppe zutreffend charakterisieren, können (vor allem bei atypischen Gruppenmitgliedern) gänzlich fehlgehen. Das ist genau das Argument, auf das auch das finnische nationale Nichtdiskriminierungs- und Gleichheitstribunal seine Verurteilung des Kreditinstituts Svea Ekonomi gestützt hat. Zwar mag es durchaus sein, dass die Mehrzahl von Männern mit finnischer Staatsangehörigkeit Probleme bei der Kreditrückzahlung haben. Über die Solvenz und Leistungsfähigkeit des konkreten Antragstellers sagt das hingegen nicht unbedingt etwas aus. Aus *sozio-ökonomischer Perspektive* wird der Generalisierungseffekt beim Einsatz von ADM-Systemen genutzt, um auf Grundlage statistischer Aussagen Informationsdefizite kostengünstig zu überwinden.[59] Das Generalisierungsrisiko beschreibt deshalb in normativer Hinsicht einen Konflikt zwischen Effizienz und (Einzelfall-) Gerechtigkeit.[60] Dieser kann nicht (allein) durch technische oder organisatorische Maßnahmen der „Fehlervermeidung" aufgelöst werden, sondern bedarf – als genuiner Wertkonflikt – einer gesellschaftlichen Abwägungsentscheidung.[61]

---

[57] Aus deutsch-privatrechtlicher Sicht paradigmatisch *Grünberger*, Personale Gleichheit. Der Grundsatz der Gleichbehandlung im Zivilrecht, 2013, 530–534. Aus der Perspektive des deutschen Verfassungsrechts (für die mittelbare Diskriminierung) *Mangold*, Demokratische Inklusion durch Recht, 2021, insbesondere 243–246. Nachweise zum U.S. Recht bei *Bagenstos*, California Law Review 2006, 1. Zum Ganzen aus theoretischer Sicht *Khaitan*, A Theory of Discrimination Law, 2015, 121 und aus rechtsökonomischer Perspektive *McAdams*, Harvard Law Review 1995, 1003.

[58] Allgemein dazu bei statistischen Differenzierungsprozessen *Britz* (Fn. 45), S. 133–136. Aus rechtlicher Perspektive im ADM-Kontext *Zarsky*, (Fn. 6), S. 1409: „imperfect generalization"; *Orwat* (Fn. 8), S. 86–89; *Xenidis/Senden* (Fn. 20), S. 158.

[59] Das ist die zentrale Einsicht der ökonomischen Modelle statistischer Diskriminierung. Siehe dazu die grundlegenden Arbeiten von *Phelps*, The American Economic Review 1972, 659 und *Arrow*, in: Ashenfelter (Hrsg.), Labor economics. Labor market discrimination, labor mobility and compensating, 1995, 3. Umfassender Überblick bei *Fang/Moro*, in: Benhabib/Jackson/Bisin (Hrsg.), Handbook of Social Economics, 2011, 133.

[60] *Britz* (Fn. 45), S. 133–136. Siehe zum Problem auch *Gandy Jr.* (Fn. 7), S. 65–67.

[61] Ähnlich *Wachter/Mittelstadt/Russel*, Computer Law & Security Review 2021, Artikel 105567, 4 et passim.

Damit wird die normative Problematik des Generalisierungseffektes deutlich. *Gabriele Britz* hat bereits früh herausgearbeitet, dass aus *normativ-rechtlicher* Perspektive das „spezifische Gleichheitsproblem statistischer Diskriminierung“[62] im „Generalisierungsunrecht“ statistischer Differenzierungsprozesse, sowie in seinem Spannungsverhältnis zur Einzelfallgerechtigkeit besteht. Auch aus freiheitsrechtlicher Perspektive ist das Generalisierungsrisiko bedenklich, weil eine unerwünschte Zuschreibung von Fremdbildern stattfindet, die den Anspruch des Individuums beeinträchtigt, selbst mitzubestimmen, wie es der Öffentlichkeit gegenübertritt.[63] Rechtsdogmatisch kann das als Beeinträchtigung des Allgemeinen Persönlichkeitsrechts rekonstruiert werden.[64] Dieser rechtlichen Beurteilung entsprechen aus *normativ-ethischer Perspektive* diejenigen deontologischen Ansätze zur Legitimierung des Nichtdiskriminierungsrechts, die den ethischen Unwert von Diskriminierung daraus ableiten, dass der einzelnen Person ihr Anspruch als ein Individuum behandelt zu werden verweigert wird.[65]

c) Korrelationsrisiko

Das *Korrelationsrisiko* beruht auf der zentralen *sozio-technischen Funktionsweise* von ADM-Systemen: Sie versuchen, bestimmte Zielvariablen anhand von Input-Parametern zu optimieren.[66] Auf diese Weise sollen im Wege des Induktionsprinzips generalisierbare Aussagen aus Einzelfällen gewonnen werden.[67] Für ADM-Systeme kommt es dafür (bislang häufig)[68] nicht entscheidend darauf an, die (tatsächlich oder vielleicht auch gerade nicht) zugrundeliegenden Kausalstrukturen zu verstehen. Darin liegt der entscheidende Unterschied zu klassischen

---

[62] *Britz* (Fn. 45), S. 134.

[63] Vgl. dazu paradigmatisch BVerfG Beschl. v. 3.6.1980 – 1 BvR 185/77 (Rn. 16, 17) = BVerfGE 54, 148: „Der Einzelne soll – ohne Beschränkung auf seine Privatsphäre – grundsätzlich selbst entscheiden können, wie er sich Dritten oder der Öffentlichkeit gegenüber darstellen will, ob und inwieweit von Dritten über seine Persönlichkeit verfügt werden kann. […] Im Zusammenhang hiermit kann es nur Sache der einzelnen Person selbst sein, über das zu bestimmen, was ihren sozialen Geltungsanspruch ausmachen soll; insoweit wird der Inhalt des allgemeinen Persönlichkeitsrechts maßgeblich durch das Selbstverständnis seines Trägers geprägt.“.

[64] Zu einem solchen Ansatz siehe *Britz* (Fn. 45), S. 179–209. Dazu aus theoretischer Perspektive *Wielsch*, JZ 2020, 105.

[65] Exemplarisch *Miller*, Principles of Social Justice, 2001, 168–169. Differenzierter dagegen *Lippert-Rasmussen*, The Journal of Ethics 2011, 47. Kritisch *Schauer*, in: Lippert-Rasmussen (Hrsg.), The Routledge Handbook of the Ethics of Discrimination, 2020, 42 (49–51).

[66] Zur Funktionsweise siehe *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen, 2018, 30–34.

[67] *Shalev-Shwartz/Ben-David* (Fn. 2), S. 1–3; *Barocas/Hardt/Narayanan* (Fn. 28), S. 1–3 und 4–6.

[68] Zur Debatte, ob und inwiefern beim maschinellen Lernen Kausalitätskonzepte einbezogen werden sollten *Schölkopf*, arXiv:1911.10500v2, Edition 23.12.2019; *Barocas/Hardt/Narayanan* (Fn. 28), S. 79–120.

Ansätzen statistischer Modellierung.[69] Zum *normativen Problem* wird dieses Charakteristikum ADM-basierter Entscheidungsfindung, weil menschliche Entscheidungs- und Rationalitätsstandards gemeinhin an ein (wie auch immer geartetes) Konzept der Kausalität anknüpfen.[70] Zwei Phänomene sind deshalb für dieses von ADM-Systemen ausgehende Korrelationsrisiko charakteristisch: die besondere Diskriminierungsform der *proxy discrimination* (bzw. besser *correlation discrimination*) einerseits[71] und andererseits der Vorwurf, ADM-Systeme würden Entscheidungen nach (augenscheinlich) irrelevanten oder irrationalen Kriterien treffen.[72]

Das in der Literatur zumeist als *proxy discrimination* bezeichnete Phänomen tritt auf, wenn ADM-Systeme im zugrundeliegenden Datenmodell an (vermeintlich) neutrale Kriterien anknüpfen und diese nicht nur mit der gesuchten Zielvariable, sondern auch mit einer durch das Nichtdiskriminierungsrecht verpönten Kategorie korrelieren.[73] Dies ist in zweifacher Hinsicht bedenklich: Zum einen, weil dadurch die Wahrscheinlichkeit einer rechtlich relevanten algorithmischen Diskriminierung steigt. Und zum anderen, weil diese Form der Diskriminierung häufig schwierig festzustellen und nachzuweisen ist, also in intransparenter Weise erfolgt (allgemein zum Transparenzrisiko siehe II.2.d.).[74] Entscheidend für die *normativ-rechtliche Bewertung* dieses Phänomens ist das bereits angesprochene Auseinanderfallen von algorithmischer Handlungslogik (Optimierung von Zielvariablen auf Grundlage von Korrelationen) und menschlichen Bewertungsmaßstäben (Rationalitätsstandards, die an Kausalitätskonzepte anknüpfen). Diesem Auseinanderfallen kann eine differenzierende Sichtweise auf das zugrundeliegende faktische Problem Rechnung tragen. Bei Lichte betrachtet beinhaltet das in der Literatur als *proxy discrimination* bezeichnete Phänomen konzeptionell nämlich zwei Unterfälle, die sich in ihrer normativen Bewertung unterscheiden: (1.) Von einer *proxy discrimination*[75] möchte ich deshalb nur dann sprechen, wenn ein ADM-System an (vermeintlich) neutrale Kriterien anknüpft, die unterschiedliche

---

[69] Grundlegend und noch immer aktuell dazu *Breiman*, Statistical Science 2001, 199.

[70] Andeutungsweise auch *Gerards/Xenidis*, Algorithmic discrimination in Europe. Challenges and opportunities for gender equality and non-discrimination law, 2021, 44. Ähnlich aus wissenschaftstheoretischer Sicht *Calude/Longo*, Foundations of Science 2017, 595.

[71] Dazu aus technischer Sicht *Calders/Žliobaitė* (Fn. 25), S. 52–53. Aus der Perspektive des europäischen Nichtdiskriminierungsrechts *Hacker* (Fn. 20) S. 1148–1150; *Xenidis/Senden* (Fn. 20), S. 155–156, 158–160, 172–175. Aus der Sicht des U.S. Nichtdiskriminierungsrechts exemplarisch *Barocas/Selbst* (Fn. 20), insbesondere S. 691–692, 701–714, 720–722; *Cofone*, Hastings Law Journal 2019, 1389; *Prince/Schwarcz*, Iowa Law Review 2020, 1257. Aus sozialwissenschaftlicher Perspektive *Williams/Brooks/Shmargad*, Journal of Information Policy 2018, 78.

[72] Allgemein aus ethischer Sicht *Halldenius*, in: Lippert-Rasmussen (Hrsg.), The Routledge Handbook of the Ethics of Discrimination, 108. Aus rechtlicher Perspektive im ADM-Kontext *Citron/Pasquale* (Fn. 45), S. 24: „arbitrariness-by-algorithm"; *Zarsky*, (Fn. 6) S. 1408–1411: „arbitrariness-by-algorithm"; *Zarsky* (Fn. 5), S. 127–128.

[73] Zur Problembeschreibung exemplarisch *Barocas/Selbst* (Fn. 20), S. 691–692.

[74] Ebenso *Hacker* (Fn. 20), S. 1149.

[75] Für eine solche enge Lesart des Begriffs auch *Prince/Schwarz* (Fn. 71).

Auswirkungen auf verschiedene Gruppen haben (wobei eine durch eine verpönte Kategorie geschützt ist) und das System seine statistische Vorhersagekraft hinsichtlich der Zielvariable *gerade aus dieser besonderen Auswirkung auf die geschützte Gruppe* schöpft.[76] (2.) Von einer *incidental discrimination* möchte ich hingegen sprechen, wenn ein ADM-System an (vermeintlich) neutrale Kriterien anknüpft, die unterschiedliche Auswirkungen auf verschiedene Gruppen haben (wobei eine durch eine verpönte Kategorie geschützt ist) und diese unterschiedliche Auswirkung *lediglich zufällig* erfolgt. Der zentrale Unterschied zwischen den beiden Konstellationen ist also, dass die besondere Auswirkung auf eine geschützte Gruppe im ersten Fall modelltreibend ist, im zweiten Fall hingegen nicht. Gemeinsam ist beiden Fällen, dass sie darauf beruhen, dass (vermeintlich) neutrale Kriterien sowohl mit der Zielvariablen wie auch mit einer geschützten Kategorie korrelieren. Als gemeinsamen Oberbegriff schlage ich deshalb den Begriff *correlation discrimination* vor. Mit dieser Unterteilung des Phänomens der *correlation discrimination* wäre es möglich, algorithmische Handlungslogik und menschliche Rationalitätsstandards zumindest teilweise konzeptionell zu synchronisieren. Die praktische Umsetzung stünde allerdings vor nicht unerheblichen Herausforderungen.[77]

Seine zweite Ausprägung findet das Korrelationsrisiko in der verbreiteten Kritik, ADM-Systeme würden Entscheidungen nach (vermeintlich) irrelevanten oder irrationalen Kriterien treffen.[78] Wenn im Fall der ADM-basierten Kreditvergabe Unternehmen beispielsweise an die Muttersprache von Antragsteller*innen anknüpfen, findet sich zumindest keine einfache oder intuitiv plausible Erklärung, warum genau dieses Kriterium für die Kreditvergabe relevant sein sollte. In solchen Fällen wird in der Literatur deshalb vielfach von der Verwendung irrelevanter Kriterien gesprochen und die Unterscheidung als irrational gebrandmarkt.[79] Umso erstaunlicher ist es deshalb, dass von Entwickler*innen und Anwender*innen von ADM-Systemen zumeist geltend gemacht wird, es liege ein Fall vollkommen rationaler Differenzierung vor. So wirbt das Recruitment-Unternehmen *JobTarget* etwa wie folgt:[80]

„Artificial intelligence (AI) is intelligence demonstrated by machines. […] A machine makes rational decisions based on algorithms that mimic cognitive functions."

„Artificial intelligence means your hiring process is optimized. Using Programmatic means your hiring decisions will be faster and better-informed. […] Targeted ads will be seen by the right candidates at the right time. In addition, it saves time because it can analyze large amounts of data in minutes saving resources and money."

---

[76] Vgl. *Grimmelmann/Westreich*, California Law Review Online 2017, 164 (170, 172); *Prince/Schwarcz* (Fn. 71), S. 1261.

[77] Vgl. nur *Pope/Syndor*, American Economic Journal 2011, 206: „Of course, both arguments have merits – these variables are likely neither solely predictive nor purely proxies for omitted characteristics.".

[78] Siehe dazu die Nachweise in Fn. 72.

[79] Siehe dazu die Nachweise in Fn. 72.

[80] Vgl. https://www.jobtarget.com/blog/how-artificial-intelligence-is-changing-recruitment/.

Ein ähnliches Bild zeichnet auch die rechtswissenschaftliche Literatur, wenn es um die Frage der Rechtfertigung der genannten Diskriminierung geht:[81] Ein legitimer Zweck liege beim Scoring in der Regel vor, um Kreditausfallrisiken zu minimieren oder geeignete Arbeitnehmer*innen ausfindig zu machen.[82] Geeignet seien ADM-Systeme zumeist auch, weil sie in effizienter Weise auf statistischer Evidenz aufbauen. Die Erforderlichkeit ergebe sich aus der Tatsache, dass algorithmische Entscheidungen häufig eine größere Genauigkeit aufweisen würden als alternative (insbesondere menschliche) Entscheidungsprozesse. Wie lässt sich diese Spannung von „irrationaler Diskriminierung" einerseits und gerechtfertigter Diskriminierung andererseits erklären? Immerhin ist es die zentrale Aufgabe der Rechtfertigungsprüfung, auf die Einhaltung von Rationalitäts- und Angemessenheitsstandards zu bestehen.[83] Die Antwort wird offensichtlich, wenn man den *sozio-ökonomischen Hintergrund* der Differenzierung betrachtet und berücksichtigt, dass die Konzepte „Relevanz" und „Rationalität" häufig eine Frage der Perspektive sind: Aus der Sicht der Antragsteller*innen liegt eine „irrationale" Differenzierung vor, weil die Bank an Kriterien anknüpft, die entweder in keiner Kausalbeziehung zum Differenzierungsziel (Ermittlung der Kreditwürdigkeit) stehen oder die nach der „Rationalität" des von ihnen zugrunde gelegten gesellschaftlichen Teilsystems (zum Beispiel Zugang zum Wohnungsmarkt per Kredit) nicht adäquat sind. Aus der Sicht der Bank als Anwenderin des ADM-Systems ist die Differenzierung schon deshalb rational, weil die Anknüpfung an reine Korrelationen bei aggregierten Größen monetäre Vorteile verspricht.[84] Das bringt auch die Werbung von *JobTarget* klar auf den Punkt. Ob die Differenzierung im Einzelfall auch tatsächlich zu einer korrekten Vorhersage führt, ist hingegen irrelevant. Nach der „Rationalität" des von der Bank zugrunde gelegten gesellschaftlichen Teilsystems (Kreditwirtschaft) ist sie deshalb vollkommen adäquat. Diese Perspektivenvielfalt wirkt sich auf die *normativ-rechtliche Bewertung* des Korrelationsrisikos aus: Die Frage nach der „Rationalität" oder „Irrationalität" algorithmischer Diskriminierung – wie sie die ganz überwiegende Literatur aufwirft – ist als solche falsch gestellt. Die Analyse zeigt, dass es nicht um die Dichotomie von „Rationalität" und „Irrationalität" geht, sondern um eine Kollision unterschiedlicher sozialer Handlungsrationalitäten (Plural). Dem Nichtdiskriminierungsrechts fällt daher die normative Aufgabe zu, diese unterschiedlichen sozialen Handlungslogiken zum Ausgleich zu bringen.[85]

---

[81] Exemplarisch dafür *Hacker* (Fn. 20), S. 1160–1165; *Xenidis/Senden* (Fn. 20) S. 173–175.

[82] Kritisch hinsichtlich der korrekten Wahl des Bezugspunkts der Prüfung *Gandy Jr.* (Fn. 45), S. 39.

[83] Dazu *Grünberger* (Fn. 57), v.a. S. 802–804: „Recht auf Rechtfertigung".

[84] Ebenso *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 16.

[85] Zu dieser Aufgabe des Nichtdiskriminierungsrechts siehe *Grünberger* (Fn. 57), insbesondere 802–804; *Wielsch*, in: Grundmann/Thiessen (Hrsg.), Von formaler zu materialer Gleichheit, 2021, 125.

d) Transparenzrisiko

Mit den Stichworten Intransparenz und Opazität ist das letzte Risiko datenbezogener Diskriminierung durch ADM-Systeme angesprochen. Der sog. *black box*-Charakter[86] dieser Systeme erschwert den Betroffenen häufig die Durchsetzung ihres Anspruchs auf Nichtdiskriminierung und schafft dadurch ein *Transparenzrisiko*.[87] Dieses hat zwei zeitlich gestufte Dimensionen: Zunächst sind die Betroffenen häufig bereits nicht in der Lage, zu erkennen, dass ein Fall (potenzieller) Diskriminierung vorliegt und verkennen deshalb die Notwendigkeit eines (gerichtlichen) Rechtsschutzes.[88] Sodann sind sie im Prozess auch selten imstande, ihren Darlegungs- und Beweisanforderungen zum Nachweis der Diskriminierung zu genügen.[89] Die *sozio-technischen und rechtlichen Hintergründe* des Transparenzrisikos sind vielfältig: Dass die betroffenen Personen eine potentielle Diskriminierung nicht erkennen, liegt zum einen daran, dass sie häufig nicht wissen, dass ein ADM-System zur Entscheidungsfindung eingesetzt wurde.[90] Zum anderen wird die Erkennbarkeit dadurch eingeschränkt, dass ADM-basierte Vertragsschlüsse höchst personalisiert stattfinden und daher (auch den Betroffenen) der für den Diskriminierungstatbestand erforderliche Vergleich mit anderen (hypothetischen) Personen schwerfällt.[91] Und selbst wenn eine potenzielle Diskriminierung als solche identifiziert wurde, ist die Nachvollziehbarkeit der Entscheidung häufig stark eingeschränkt. Diese Einschränkung kann sich aus rechtlichen wie auch aus rein faktischen Gründen ergeben:[92] Rechtliche Hindernisse bestehen, wenn Anwender*innen Einblicke in die Funktionsweise des Systems mit Hilfe des Immaterialgüterrechts, des Geschäftsgeheimnisschutzes oder des Datenschutzrechts abwehren können.[93] Faktisch ist die Nachvollziehbarkeit eingeschränkt, weil es bei bestimmten Methoden des maschinellen Lernens nach

---

[86] Begriffsprägend *Pasquale* (Fn. 14).

[87] Instruktiv und mit Beispielen dazu *O'Neil* (Fn. 10), S. 10, 28–29; siehe auch *Gandy Jr.* (Fn. 45), S. 40. Aus technischer Perspektive grundlegend *Burrell*, Big Data & Society 2016, 1. Den technischen Diskurs zusammenfassend *Das/Rad*, arXiv:2006.11371v2, Edition 23.6.2020. Aus rechtlicher Perspektive im ADM-Kontext *Kroll* et al., University of Pennsylvania Law Review 2017, 633; *Hacker* (Fn. 20), S. 1167–1170; *Olsen* et al., iCourts Working Paper Series 2019, No. 162; *Xenidis/Senden* (Fn. 20), S. 175–181; *Gerards/Xenidis* (Fn. 70), S. 45–46 und 75; *Grünberger*, Reformbedarf im AGG: Beweislastverteilung beim Einsatz von KI, ZRP 2021, 232. Dezidiert positive Einschätzung bei *Kleinberg* et al. (Fn. 20).

[88] Instruktiv dazu *O'Neil* (Fn. 10), S. 28–29. Ebenso *Hacker* (Fn. 20), S. 1169.

[89] Siehe dazu die Nachweise in Fn. 87.

[90] Aus diesem Grund plädiert insbesondere *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, 2019, 341 für die Einführung von Informationspflichten über den Einsatz von ADM-Systemen. Art. 22 DSG-VO hilft hier nur selten weiter, weil er nach überwiegender Lesart nur eingreift, wenn vollautomatisiert entschieden wird.

[91] Ähnlich *Mann/Matzner*, Big Data & Society Vol. 6 No. 2 (2019), 1 (4).

[92] Siehe zu den folgenden drei Punkten den systematisierenden Überblick bei *Burrell* (Fn. 87) m. w. N.

[93] Dazu *Martini* (Fn. 90) m. w. N.

heutigem Stand von Wissenschaft und Technik schwierig ist, anzugeben, welchen Anteil einzelne Inputfaktoren am Output des Systems hatten. Aber auch für Expert*innen leicht nachvollziehbare Systeme sind häufig in komplexe, arbeitsteilige Mensch-Maschine-Interaktionen eingebunden, die für Laien nur schwer zu durchschauen sind. Ebenso vielfältig wie diese *sozio-technischen und rechtlichen Hintergründe* des Transparenzrisikos sind seine *sozio-ökonomischen Ursprünge:* Zum einen bedingen sich Präzision und fehlende Nachvollziehbarkeit von ADM-Systemen häufig gegenseitig.[94] Die Intransparenz ist dann zwar nicht beabsichtigt, sie wird aber als Korrelat der gesteigerten Vorhersagekraft des ADM-Systems in Kauf genommen. Die eingeschränkte Nachvollziehbarkeit kann aber auch als solche vorteilhaft sein. Dies ist insbesondere der Fall, wenn durch die Intransparenz anderweitige wirtschaftliche Interessen wie Geschäftsgeheimnisse geschützt werden sollen oder Anwender*innen ein *gaming-the-system* durch die Nutzer*innen verhindern wollen.

Die *normativ-rechtliche Problematik* des Transparenzrisikos besteht darin, dass die Intransparenz des Entscheidungsprozesses die effektive Rechtsdurchsetzung des Nichtdiskriminierungsrechts beeinträchtigt.[95] Die Betroffenen können entweder nicht erkennen, dass ihr Recht auf Nichtdiskriminierung verletzt wurde und (gerichtlicher) Rechtsschutz erforderlich ist oder es stehen ihnen nicht die Mittel zur Verfügung, den Rechtsweg mit hinreichender Aussicht auf Erfolg beschreiten zu können. Dieses Risiko wird den von Diskriminierung Betroffenen beim Einsatz von ADM-Systemen faktisch vollständig zugewiesen. Das Transparenzrisiko ist deshalb ein faktisch-normativer Begriff:[96] Faktisch, weil die tatsächliche Möglichkeit des Erkennens von Diskriminierung und der Nachvollziehbarkeit entsprechender Systeme eingeschränkt ist. Normativ, weil es Aufgabe des Nichtdiskriminierungsrechts ist, eine Antwort auf die Frage zu finden, ob diese Zuweisung auch rechtlich-normativ abgesichert werden soll oder nicht.

3. Zwischenfazit

Das Spezifikum ADM-basierter Diskriminierung ist ihre Datenbezogenheit. Sie kann dem algorithmischen Datenmodell entspringen (*biased training data*), den abgebildeten gesellschaftlichen Realitäten entstammen (*unequal base rates*), auf Eingriffen in das Datenmodell beruhen (*bias in modeling*) oder aber das Resultat einer fehlerhaften Anwendung des Systems sein (*bias in usage*). Charakterisiert wird diese algorithmische, datenbezogene Diskriminierung durch vier Risiken: das Perpetuierungsrisiko, das Generalisierungsrisiko, das Korrelationsrisiko und

---

[94] Dazu aus technischer Perspektive *Kamwa/Samantaray/Joos*, IEEE Transactions on Smart Grid 2012, 152.

[95] Dazu *Grünberger* (Fn. 87).

[96] *Grünberger* (Fn. 87).

das Transparenzrisiko. In einem zweiten Schritt zeichne ich nun kurz nach, wie das Nichtdiskriminierungsrecht diesen Risiken bereits heute begegnen kann und welche dogmatischen Hürden dabei auftreten (können).

III. Gegenwärtige dogmatische Herausforderungen

Versucht man, die paradigmatischen Fälle ADM-basierter Diskriminierung in die bestehende Dogmatik des Nichtdiskriminierungsrechts einzuordnen, stellen sich Fragen auf allen „Ebenen" des Rechtsgebiets. Die gegenwärtigen Herausforderungen betreffen Anwendungsbereich, Tatbestand, sowie Rechtfertigungsprüfung und reichen bis hin zu den Rechtsfolgen und Fragen der Rechtsdurchsetzung.

1. Anwendungsbereich des Nichtdiskriminierungsrechts

Das europäische und deutsche Nichtdiskriminierungsrecht enthält ein *beschäftigungsrechtliches*[97] (III.1.a.) und ein *allgemein-zivilrechtliches Benachteiligungsverbot*[98] (III.1.b.). Letzteres erstreckt die Reichweite des Nichtdiskriminierungsgrundsatzes ins allgemeine Vertragsrecht. Der Einsatz von ADM-Systemen ist in beiden Bereichen bereits heute Realität.[99] Die Frage, die im Folgenden beantwortet werden soll, ist, ob der Anwendungsbereich des Nichtdiskriminierungsrechts für den Einsatz von ADM-Systemen adäquat ausgestaltet ist. Ich möchte zeigen, dass ADM-basierte Diskriminierung im Beschäftigungssektor insoweit keine oder kaum Besonderheiten mit sich bringt, während sie im Anwendungsbereich des allgemeinen zivilrechtlichen Benachteiligungsverbots dafür umso mehr Herausforderungen aufwirft.

a) Einsatz von ADM-Systemen im Beschäftigungssektor

Das beschäftigungsrechtliche Benachteiligungsverbot erfasst die in der Literatur diskutierten Fälle ADM-basierter Diskriminierung in der Regel ohne größere Probleme. Das beruht vor allem auf zwei Gründen: (1.) Erstens ist der sachliche Anwendungsbereich des beschäftigungsrechtlichen Diskriminierungsverbots sehr breit ausgestaltet. Erfasst ist „die gesamte Zeitspanne des Vertragsverhältnisses […], von der Vertragsanbahnung über die Vertragsdurchführung bis zur Vertragsbeendigung und -abwicklung."[100] Der Nichtdiskriminierungsgrundsatz beansprucht damit insbesondere Geltung für die intensiv diskutierten Fälle diskriminierender Einstellungsverfahren, die das Stadium der Vertragsanbahnung

---

[97] Vgl. RL-2000/43/EG, RL-2000/78/EG, RL-2006/54/EG, § 2 Abs. 1 Nr. 1–4 AGG.
[98] Vgl. RL-2000/43/EG, RL-2004/113/EG, § 2 Abs. 1 Nr. 5–8 AGG.
[99] Siehe dazu die umfangreichen Nachweise bei *Orwat* (Fn. 8), S. 34–75.
[100] *Grünberger* (Fn. 57), S. 601.

betreffen.[101] Auch die Konstellation eines (potentiellen) *digital gender pay gap*[102] ist erfasst, weil es um die Frage nach der Entgeltgleichheit männlicher und weiblicher Beschäftigter geht, die seit jeher Kernbestandteil des beschäftigungsrechtlichen Diskriminierungsschutzes ist.[103] (2.) Problematischer sind dagegen die Fälle, die den persönlichen Anwendungsbereich des beschäftigungsrechtlichen Diskriminierungsverbots betreffen. Hier ist in der Regel die Arbeitnehmereigenschaft der betroffenen Personen fraglich. Paradigmatisch dafür steht die Diskussion, ob die Fahrer*innen des Plattformdienstleisters *Uber* Arbeitnehmer im rechtlichen Sinne sind und damit durch das Diskriminierungsverbot geschützt werden.[104] Diese Frage mag im Ergebnis schwierig zu beantworten sein. Im Kern geht es dabei aber nicht um ein Problem, das seine Ursache in der sozio-technischen Funktionalität von ADM-Systemen hat. Es geht vielmehr um veränderte Formen sozialer Machtstrukturen im modernen Arbeitsleben der Plattformökonomie.[105] Das ist der zweite Grund, warum der Einsatz von ADM-Systemen *als solcher* keine großen Besonderheiten für das beschäftigungsrechtliche Diskriminierungsverbot mit sich bringt.

b) Einsatz von ADM-Systemen im allgemeinen Zivilrecht

Anders verhält es sich beim allgemein-zivilrechtlichen Benachteiligungsverbot. Hier hat *Philipp Hacker* darauf aufmerksam gemacht, dass eine konservative Interpretation des sachlichen und persönlichen Anwendungsbereichs dem Nichtdiskriminierungsgrundsatz im allgemeinen Vertragsrecht sämtliche Zähne ziehen könnte.[106] Eine Auslegung, die historische Begriffsverständnisse kontextualisiert und technik-responsiv[107] auf die spezifische Eigenart algorithmischer Diskriminierung Rücksicht nimmt, kann dem Nichtdiskriminierungsrecht hingegen bereits heute in den meistdiskutierten Fällen zur Anwendbarkeit verhel-

---

[101] Vgl. aus der rechtswissenschaftlichen Literatur zu diesen Fällen *Dzida/Groh*, NJW 2018, 1917; *von Lewinski/de Barros Fritz*, NZA 2018, 620; *Steege*, MMR 2019, 715; *Freyler*, NZA 2020, 284; *Kullmann*, in: Beyer et. al. (Hrsg.), Privatrecht 2050 – Blick in die digitale Zukunft, 2020, 227; *Xenidis/Senden* (Fn. 20) S. 163–167; *Wimmer*, Algorithmusbasierte Entscheidungsfindung als Methode des diskriminierungsfreien Recruitings, 2022, S. 322–323.

[102] Dazu *Xenidis/Senden* (Fn. 20) S. 161–162; *Gerards/Xenidis* (Fn. 70), S. 55–56.

[103] Vgl. Art. 157 AEUV, Art. 1 lit. b und Art. 4 RL-2006/54/EG und § 2 Abs. 1 Nr. 2 AGG.

[104] *Xenidis/Senden* (Fn. 20), S. 161–162; *Gerards/Xenidis* (Fn. 70), S. 56–57.

[105] Umfassend zur Problemstellung und zur internationalen Diskussion mit zahlreichen Nachweisen *Leist/Hießl/Schlachter*, Plattformökonomie – eine Literaturauswertung (Forschungsbericht / Bundesministerium für Arbeit und Soziales, FB499), 2017, S. 32–61.

[106] *Hacker* (Fn. 20), S. 1154–1160.

[107] Vgl. zu diesem Ansatz aus der Sicht des Technikrechts *Sommer*, Haftung für autonome Systeme, 2020, S. 55–65. Allgemein aus methodischer Perspektive *Grünberger*, AcP 219 (2019), 924; kritisch dazu *Riesenhuber*, AcP 219 (2019), 892. Grundlegend *Nonet/Selznick*, Toward Responsive Law – Law & Society in Transition, 2017 (Nachdruck des Originals von 1978). Aus rechtstheoretischer Perspektive *Luhmann*, Das Recht der Gesellschaft, 7. Auflage 2018 (Nachdruck des Originals von 1995), S. 225–226.

fen. Paradigmatisch für diese Hypothese stehen insbesondere vier Problemkreise: (1.) Zu ihnen gehört zunächst die Auslegung des Begriffs der „Dienstleistung" im AGG bzw. den Nichtdiskriminierungsrichtlinien. Klassisch beinhaltet dieser nämlich ein Entgeltlichkeitserfordernis im Sinne einer monetären Kompensation.[108] Das wirft die Frage auf, ob ADM-basierte Dienste, die nach dem Motto „Bezahlen mit Daten" funktionieren (wie etwa automatische Übersetzungs- oder online Gesichtserkennungssoftware)[109] überhaupt vom Begriff der Dienstleistung erfasst sind. Wie *Philipp Hacker* gezeigt hat, kann eine historisch-kontextualisierende Auslegung hier weiterhelfen:[110] Berücksichtigt man nämlich, dass das Entgeltlichkeitserfordernis des europäischen Dienstleistungsbegriffs noch aus einer Zeit stammt, in der das Phänomen des „Bezahlens mit Daten" nicht bekannt war[111] und dass dieses Erfordernis primär dem Zweck dient, die Anwendbarkeit des Diskriminierungsverbotes auf professionelle Tätigkeiten zu beschränken, wird man kaum umhin kommen, auch eine Kompensation durch Überlassung von Daten genügen zu lassen. (2.) Der zweite Problemkreis betrifft die Frage der Anwendbarkeit des Benachteiligungsverbots auf stereotypisierende Medien (insbesondere Suchmaschinen) und Werbung.[112] Das Problem ist auf sozialwissenschaftlich-technischer Seite gut erforscht.[113] Ob die entsprechenden Fälle vom Nichtdiskriminierungsrecht erfasst werden, ist hingegen alles andere als gewiss. Hinsichtlich der verpönten Kategorie ‚Geschlecht' macht RL-2004/113/EG nämlich eine explizite Ausnahme für den „Inhalt von Medien und Werbung".[114] Eine Interpretation, die diese Ausnahme allein auf den *Inhalt* von Medien und Werbung beschränkt und die *Verbreitung* und *Wirkung* vom Nichtdiskriminierungsverbot erfasst sehen will, ist zwar möglich,[115] besonders naheliegend erscheint sie mir indes nicht.[116] Für die verpönten Kategorien der „Rasse" und der ‚ethnischen Herkunft' fehlt es

---

[108] Der unionale Begriff der Dienstleistung verlangt klassisch Entgeltlichkeit. Für das IPR/IZVR siehe EuGH Urt. v. 23.4.2009 – C-533/07, Slg. 2009, I-03327 – Falco Privatstiftung; für das Primärrecht siehe EuGH Urt. v. 12.12.1974 – 36/74, Slg. 1974, 01405 – Walrave und Koch. Anders aber ErwGr (17) der Richtlinie 2000/31/EG: „in der Regel gegen Entgelt".

[109] Beispiele nach *Hacker* (Fn. 20) S. 1155.

[110] *Hacker* (Fn. 20) S. 1155–1156.

[111] Anders jetzt aber Art. 3 Abs. 1, S. 2 RL-2019/770/EU.

[112] Zu diesem Problem *Gerards*/*Xenidis*, (Fn. 70), S. 58–60.

[113] Zum Problem des *bias* bei Suchmaschinen siehe *Noble* (Fn. 42); *Kay*/*Matuszek*/*Munson*, CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, 3819. Zum Problem des *bias* in der Werbung *Sweeney*, ACM Queue 2013, 10; *Lambrecht*/*Tucker*, Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads, Edition 5.9.2021, abrufbar unter https://ssrn.com/abstract=2852260; *Ali* et al., Proceedings of the ACM on Human-Computer Interaction, 2019, Article 199.

[114] Art. 3 Abs. 3 RL-2004/113/EG.

[115] Zu diesem Gedanken *Gerards*/*Xenidis* (Fn. 70), S. 60.

[116] Die Europäische Kommission scheint das ähnlich zu sehen, vgl. Leitlinien zur Anwendung der Richtlinie 2004/113/EG des Rates auf das Versicherungswesen im Anschluss an das Urteil des Gerichtshofs der Europäischen Union in der Rechtssache C-236/09 (Test-Achats), KOM(2011)9497 endg.

dagegen an einer ausdrücklichen Regelung.[117] Ein Umkehrschluss zur expliziten Ausnahme des Art. 3 Abs. 3 RL-2004/113/EG liegt damit zwar nahe,[118] gerichtlich entschieden ist dies bislang jedoch nicht. Auch im AGG ist die Frage nicht explizit geregelt. Erforderlich wäre hier eine Subsumtion unter die allgemeinen Tatbestandsvoraussetzungen von § 2 AGG und § 19 AGG. Unter § 2 AGG ließen sich Werbung und Suchmaschinen wohl nur fassen, wenn man diese als Bedingungen des *Zugangs* zu Gütern und Dienstleistungen ansieht. Im Fall des § 19 AGG müsste der Begriff zivilrechtlicher Schuldverhältnisse weit, und das Merkmal „typischerweise ohne Ansehen" technik-responsiv (dazu unten) ausgelegt werden. In jedem Fall bedürfte der Anwendungsbereich des allgemein-zivilrechtlichen Benachteiligungsverbots einer expansiven Interpretation, um stereotypisierende Werbung und Suchmaschinen erfassen zu können. (3.) Gegenstand des dritten Problemkreises ist der persönliche Anwendungsbereich des zivilrechtlichen Diskriminierungsverbots. Hier geht die bislang wohl herrschende Meinung (der deutschen Literatur)[119] allein von einer Bindung der Angebotsseite, nicht aber von einer Bindung der Nachfrageseite aus.[120] Dahinter steht der Gedanke, Verbraucher*innen eine Freiheit in ihrer privaten Sphäre zu sichern.[121] Auch wenn man diese Erwägung im ‚analogen Kontext' noch für überzeugend halten mag,[122] wird sie in der ‚digitalen Welt' zumindest zweifelhaft: Im klassisch-analogen Beispiel[123] der Einkaufsentscheidung einer Verbraucher*in bei einer Bäckerei handelt es sich ersichtlich um die rein privatautonome Ausübung von Kund*innenpräferenzen im privaten Bereich. Sollte sich ein von *Philipp Hacker* in Aussicht gestelltes Szenario allerdings realisieren und zukünftig vermehrt ADM-basierte (‚smarte') IoT-Produkte für (durchaus auch gewerblich agierende) Nachfrager*innen selbstständig Einkaufsentscheidungen treffen,[124] wird man eine Frage neu beantworten müssen: Handelt es sich hierbei noch immer um eine rein privatautonome Ausübung von Präferenzen im privaten Bereich? Oder liegt in der Entscheidungsdelegation an einen dritten Akteur nicht vielmehr eine Öffnung der Transaktion in die öffentliche Sphäre und damit in den Anwendungsbereich des Nichtdiskriminierungsrechts? – Mit anderen Worten: Trägt die Prämisse der herrschenden Auffassung auch noch im digitalen Kontext? (4.) Am spannendsten dürfte wohl die Frage sein, wie sich die zunehmende Personalisierung von Vertragsabschlussmöglich-

---

[117] Siehe dazu RL-2000/43/EG.

[118] *Gerards*/*Xenidis* (Fn. 70), S. 60.

[119] Das Problem wird soweit ersichtlich primär in der deutschen rechtsdogmatischen Literatur erörtert.

[120] Paradigmatisch MüKo-BGB/*Thüsing*, 9. Auflage, § 19 AGG Rn. 126–127; BeckOGK-AGG/*Mörsdorf*, Edition 1.9.2021, § 19 Rn. 71; Ermann-BGB/*Armbrüster*, 16. Auflage, § 19 AGG Rn. 11.

[121] Vgl. MüKo-BGB/*Thüsing* (Fn. 120), Rn. 125.

[122] Berechtigte Kritik hingegen bei *Grünberger* (Fn. 57), S. 626–631. Kritisch auch Staudinger/*Serr*, Neubearbeitung 2020, § 19 AGG Rn. 52–58.

[123] Vgl. MüKo-BGB/*Thüsing* (Fn. 120), Rn. 125.

[124] *Hacker* (Fn. 20) S. 1158.

keiten und Vertragsinhalten auf die Anwendbarkeit des zivilrechtlichen Benachteiligungsverbots auswirkt.[125] ADM-Systeme ermöglichen es, Verträge in nie da gewesener Weise auf die jeweiligen Vertragspartner*innen zuzuschneiden. Und zwar sowohl hinsichtlich ihrer Person wie auch hinsichtlich der inhaltlichen Konditionen. In der Sprache des AGG ist rechtsdogmatisch damit die Frage angesprochen, ob solch personalisierte Verträge Güter und Dienstleistungen betreffen, die „der Öffentlichkeit zur Verfügung stehen"[126] und die „typischerweise ohne Ansehen der Person zu vergleichbaren Bedingungen in einer Vielzahl von Fällen zustande kommen […] oder bei denen das Ansehen der Person […] eine nachrangige Bedeutung hat und die zu vergleichbaren Bedingungen in einer Vielzahl von Fällen zustande kommen"[127]. Die jüngere Rechtsprechung des Bundesgerichtshofs zur Anwendbarkeit des AGG in ‚analogen Kontexten' lässt für die ‚digitale Zukunft' insofern nichts Gutes erahnen:

„Enthält die Prüfung des Vertragsschlusses ein stark individualisiertes, personales Element, verzichtet das Gesetz im Rahmen des § 19 I Nr. 1 AGG zugunsten der persönlichen Willensbildung des Anbieters auf eine Benachteiligungskontrolle."[128]

Geht man mit dem Bundesgerichtshof davon aus, dass eine Personalisierung von Vertragsabschluss und Vertragsinhalt den o. g. Kriterien aus § 2 AGG und § 19 AGG entgegensteht, verschließt man die Tür zum AGG in vielen Fällen, in denen ADM-Systeme eingesetzt werden (können). Das Ergebnis ist paradox: Je stärker die Personalisierung bzw. Individualisierung des Vertragsverhältnisses durch den Einsatz von ADM-Systemen ausgestaltet wird, desto geringer würde der Schutz vor den diskriminierenden Folgen dieser Personalisierung ausfallen. Das Problem ist freilich seit langem bekannt.[129] Es im ADM-Kontext dogmatisch einzuhegen, könnte mit einer Auslegung gelingen, die den technischen Besonderheiten ADM-basierter Entscheidungen Rechnung trägt und anerkennt, dass Individualisierung und Personalisierung in diesen Fällen eine andere Bedeutung haben als im ‚analogen Kontext'. Möglich erscheint das durch eine Auslegung, die es für den erforderlichen Öffentlichkeitsbezug genügen lässt, dass Güter und Dienstleistungen *ihrer Art nach* der Öffentlichkeit zur Verfügung gestellt werden, auch wenn die spezifischen Konditionen im Einzelfall variieren.[130] Anhaltspunkte für eine solche Interpretation gibt der Wortlaut des § 19 Abs. 1 Nr. 1 AGG allemal:

---

[125] Siehe dazu *Hacker* (Fn. 20) S. 1156–1157; *Gerards/Xenidis* (Fn. 70), S. 61.

[126] Vgl. Art. 3 Abs. 1 lit. h RL-200/54/EG, Art. 3 Abs. 1 RL-2004/113/EG, § 2 Abs. 1 Nr. 8 AGG.

[127] Vgl. § 19 Abs. 1 Nr. 1 AGG.

[128] BGH Urt. v. 25.4.2019 – I ZR 272/15, Rn. 18 = NJW 2020, 852; BGH Urt. v. 5.5.2021 – VII ZR 78/20, Rn. 22 = NJW 2021, 2514. Berechtigte Kritik zu diesem Aspekt der zweiten Entscheidung bei *Grünberger*, NJW 2021, 2514; *Armbrüster*, JZ 2021, 1011; *Kainer*, LMK 2021, 813516.

[129] *Grünberger* (Fn. 57), S. 609–614 mit umfassenden Nachweisen. Zum Problem siehe auch MüKo-BGB/*Thüsing* (Fn. 120), Rn. 17.

[130] *Hacker* (Fn. 20) S. 1157.

Wie bei einem besonderen Gleichheitssatz zu erwarten,[131] stellt er darauf ab, dass die verschiedenen Geschäfte zu *vergleichbaren* Bedingungen zustande kommen. *Identische* Bedingungen werden gerade nicht verlangt. *De lege ferenda* ehrlicher und konsequent dürfte es freilich sein, das Phänomen ADM-basierter Diskriminierung zum Anlass zu nehmen, um sich vom überflüssigen Tatbestandsmerkmal des „typischerweise ohne Ansehen der Person […]" zu verabschieden.[132]

2. Tatbestand und Rechtfertigung

Viele Fälle der Diskriminierung durch ADM-Systeme können bereits heute mit dem Nichtdiskriminierungsrecht erfasst werden. Im Beschäftigungssektor, weil der Einsatz der entsprechenden Systeme hier keine oder kaum besondere Fragen aufwirft, und im allgemeinen Vertragsrecht, wenn man die entsprechenden dogmatischen Begrifflichkeiten kontextualisiert und technik-responsiv interpretiert. Ist die Tür zum Nichtdiskriminierungsrecht geöffnet, stehen Probleme des Diskriminierungstatbestands (III.2.a.) und der Rechtfertigungsprüfung (III.2.b.) im Raum. Auch hier wirft der Einsatz von ADM-Systemen einige Fragen auf.

a) Diskriminierungstatbestand

Zu den drängendsten gehören auf tatbestandlicher Ebene vor allem die folgenden zwei Fragen: Erstens, wie lassen sich die rechtlichen Kategorien der unmittelbaren und mittelbaren Diskriminierung im Fall algorithmischer Diskriminierung voneinander abgrenzen (dazu unten IV.2.)? Zweitens, anhand welcher Kriterien kann und soll festgestellt werden, wann ein ADM-System mittelbar diskriminiert?[133] Diese zweite Frage bereitet in mehrfacher Hinsicht Probleme: Wie bestimmt sich die Zusammensetzung der benachteiligten Gruppe und der Vergleichsgruppe? Welche Anforderungen (insbesondere statistischer Natur) sind an einen „besonderen Nachteil" zu stellen? Die größte Schwierigkeit dürfte nicht darin liegen, diese Fragen im gerichtlichen Verfahren zu klären. Der EuGH bemüht sich insoweit im Interesse der Einzelfallgerechtigkeit um fall- und kontextbezogene Antworten.[134] Problematisch ist vielmehr, dass ADM-Systeme ihren Zweck – Informationsdefizite in einer Vielzahl von Fällen kostengünstig und rechtssicher zu überwinden – nur dann erfüllen können, wenn Entwickler*innen das Verbot der mittelbaren Diskriminierung im Großen und Ganzen bereits beim Design des ADM-Systems implementieren können. In jüngerer Zeit sind deshalb Versuche unternommen

---

[131] Vgl. zum Verhältnis von Vergleichbarkeit und Identität *Windelband*, Über Gleichheit und Identität, 1910. Weitere Nachweise zu diesem Verhältnis bei *Nef*, Gleichheit und Gerechtigkeit, 1941, 3.

[132] Ähnlich schon *Grünberger* (Fn. 57), S. 613.

[133] Dazu *Wachter/Mittelstadt/Russel* (Fn. 61).

[134] Das haben *Wachter/Mittelstadt/Russel* (Fn. 61), S. 6–18 deutlich herausgearbeitet.

worden, technische Fairness-Metriken mit den rechtlichen Vorgaben des Nichtdiskriminierungsrechts abzugleichen und so den technischen und den juristischen Diskurs zusammenzuführen.[135] *Sandra Wachter* et al. gehen beispielsweise von einem gruppenbezogenen Konzept der mittelbaren Diskriminierung aus und identifizieren einen „*gold standard*" des EuGH bei der Beantwortung der Frage, ob eine geschützte Gruppe wesentlich stärker von einer neutralen Maßnahme betroffen ist, als die Vergleichsgruppe.[136] Dieser „*gold standard*" habe im technischen Diskurs zu *Fairness in Machine Learning*[137] ein Äquivalent in Gestalt der Metrik der *conditional demographic disparity*.[138] Diese sei deshalb auch bei der rechtlichen Bestimmung des Diskriminierungstatbestands heranzuziehen. Richtig daran ist, dass ein erheblicher interdisziplinärer Forschungsbedarf besteht und technischer und juristischer Diskurs voneinander lernen können und sollten. Nicht überzeugen kann mich aber die vorgeschlagene *one size fits all*-Lösung: Die von *Wachter* et al. angeführten Nachweise aus der Rechtsprechung des EuGH betreffen allesamt Fälle der Diskriminierung nach Geschlecht oder nach „Rasse" bzw. ethnischer Herkunft. Der von ihnen identifizierte sog. „*gold standard*" der Rechtsprechung adressiert deshalb primär das Problem von *unequal base rates* (dazu oben II.1.). Um die sonstigen Ursachen diskriminierender ADM-Systeme zu erfassen, wären andere Ansätze der technischen Debatte zur *Fairness in Machine Learning* besser geeignet. Hier wird es Aufgabe der (auch rechtswissenschaftlichen) Forschung sein, Kriterien zu entwickeln, nach denen bestimmt werden kann, unter welchen Umständen die im technischen Diskurs vorhandenen Fairnesskonzepte auch mit den Anforderungen des Nichtdiskriminierungsrechts kompatibel sind. Hierbei kann und soll die oben vorgeschlagene Differenzierung hinsichtlich der Ursachen algorithmischer Diskriminierung helfen.

b) Rechtfertigungsprüfung

Auf Ebene der Rechtfertigungsprüfung geht es um zwei Gesichtspunkte: (1.) Welche Maßstäbe sind bei der Rechtfertigung ADM-basierter Diskriminierung anzulegen?[139] Hier gilt es insbesondere nach den verschiedenen Entstehungsgründen

---

[135] *Hauer/Kevekordes/Haeri*, Computer Law & Security Review 2021, Artikel 105583; *Wachter/Mittelstadt/Russel* (Fn. 61); *Wachter/Mittelstadt/Russel*, West Virginia Law Review 2021, 735.

[136] *Wachter/Mittelstadt/Russel* (Fn. 61), S. 16–18.

[137] Vgl. dazu die Metaanalysen von *Chouldechova/Roth*, arXiv:1810.08810v1, Edition 20.10.2018; *Mehrabi* et al., arXiv:1908.09635v2, Edition 17.9.2019; *Caton/Haas*, arXiv:2010.04053v1, Edition 4.10.2020. Aus ethischer Sicht *Binns*, Proceedings of Macine Learning Research 2018, 149.

[138] Grundlegend dazu *Kamiran/Žliobaitė/Calders*, Knowledge and Information Systems 2013, 613.

[139] Dazu *Hacker* (Fn. 20) S. 1160–1167; *Xenidis/Senden* (Fn. 20), S. 170–175; *Gerards/Xenidis* (Fn. 70), S. 67–73; *von Ungern-Sternberg*, in: Mangold/Payandeh (Hrsg.), Handbuch Antidiskriminierungsrecht (im Erscheinen), Vorabveröffentlichung unter https://ssrn.com/abstract=3828696, dort S. 35–38. Aus der Perspektive des U.S. Nichtdiskriminierungsrechts exemplarisch *Barocas/Selbst* (Fn. 20), S. 701–712; *Kim*, William & Mary Law Review 2017, 857 (920–923).

algorithmischer Diskriminierung zu differenzieren (dazu bereits oben II.1.). Diese Unterschiede muss ein technik-responsives Nichtdiskriminierungsrecht aufgreifen und bei der Entwicklung normativer Leitlinien zugrunde legen.[140] (2.) Die zweite Frage betrifft erneut das Verhältnis von technischem und juristischem Diskurs über diskriminierungsfreie ADM-Systeme. Konkret: In welchem Verhältnis stehen die technischen Ansätze zu *equality by design* und *Fairness in Machine Learning* zur dogmatischen Kategorie der Positivmaßnahmen?[141] Oder mit anderen Worten: Welcher Rechtfertigungsmaßstab ist im Fall eines *bias in modeling* (dazu oben II.1.) anzulegen? Handelt es sich um Positivmaßnahmen nach § 5 AGG oder nur um „ganz gewöhnliche" Benachteiligungen mit Rechtfertigungsmöglichkeiten nach §§ 8–10 AGG und § 20 AGG? Beide Problemkreise lassen sich bereits heute und innerhalb des bestehenden dogmatischen Rahmens lösen. Erforderlich ist dafür allerdings eine Rechtsdogmatik, die sich technischen und sozialwissenschaftlichen Diskursen nicht verschließt, sondern diese aufnimmt und die dort gewonnenen Erkenntnisse in die Sprache des Rechts übersetzt.

3. Rechtsfolgen und Rechtsdurchsetzung

Was die Rechtsfolgen eines Verstoßes gegen das Benachteiligungsverbot anbelangt, so steht vor allem die Frage nach dem haftenden Akteur[142] und die damit zusammenhängende Problematik der Zurechnung[143] im Zentrum der Debatte. Beides sind mittlerweile „klassische" Fragen des KI-Haftungsrechts, die auch im breiteren Diskurs zur KI-Regulierung intensiv diskutiert werden.[144] Als potenzielle Haftungssubjekte kommen vorliegend eine ganze Vielzahl an Personen in Betracht: Zunächst ist an die Händler*innen von KI-Trainingsdaten bzw. die Hersteller*innen von IoT-Produkten, die KI-Trainingsdaten erzeugen, zu denken. Sodann aber auch an die Entwickler*innen der ADM-Datenmodelle. Zuletzt sind die Anwender*innen, sowie die (mit einer (Teil-)Rechtsfähigkeit ausgestatteten)[145] ADM-Systeme selbst potenzielle Adressaten des Benachteiligungsverbotes. Die primäre Verantwortlichkeit sollte bei den Anwender*innen der Systeme gesucht werden.[146] Die deutsche Literatur stellt sich aus diesem Grund die Frage, ob und wie ein etwaiges Fehlverhalten der sonstigen genannten Personen den Anwender*innen (analog) § 278 BGB zugerechnet werden kann.[147] Sekundär ist

---

[140] Ebenso implizit auch *Hacker* (Fn. 20) S. 1160–1165, allerdings basierend auf seiner zweiwertigen Unterscheidung der Ursachen algorithmischer Diskriminierung.

[141] Andeutungsweise auch *Hacker* (Fn. 20) S. 1180–1181.

[142] Dazu *Gerards/Xenidis* (Fn. 70), S. 73–74.

[143] Dazu *von Lewinski/de Barros Fritz* (Fn. 101), S. 623; *Freyler* (Fn. 101), S. 288–290.

[144] Paradigmatisch dafür *Zech*, ZfPW 2019, 198; *Wagner*, VersR 2020, 717; *Sommer* (Fn. 107).

[145] Zur Debatte instruktiv *Schirmer*, JZ 2016, 66; *Teubner*, AcP 218 (2018), 155.

[146] So implizit wohl auch *von Lewinski/de Barros Fritz* (Fn. 101); *Dzida/Groh* (Fn. 101); *Freyler* (Fn. 101).

[147] *von Lewinski/de Barros Fritz* (Fn. 101), S. 623; *Freyler* (Fn. 101), S. 288–290.

aber auch an die selbstständige Verantwortlichkeit der sonstigen Beteiligten zu denken. *Janneke Gerards und Raphaële Xenidis* haben diesbezüglich den innovativen Vorschlag gemacht, die bislang wenig beachtete dogmatische Figur der ‚Anweisung zur Diskriminierung' zu operationalisieren.[148] Die Einzelheiten einer solchen Verantwortlichkeit müssten freilich noch ausbuchstabiert werden.

Das Problem der Rechtsdurchsetzung betrifft das Transparenzrisiko ADM-basierter Diskriminierung (dazu oben II.2.d.). Unbestritten ist mittlerweile, dass das gegenwärtige System des Individualrechtsschutzes in seiner konkreten Ausgestaltung (Stichwort: Beweislastverteilung) die Betroffenen algorithmischer Diskriminierung vor unüberwindbare Hürden stellt.[149] Die Vorschläge zur Lösung dieses Problems sind mannigfaltig. Manche wollen Kollektivverfahren und behördliche Rechtsdurchsetzung neben dem Individualrechtsschutz stärken.[150] Andere wollen unmittelbar beim Individualrechtsschutz ansetzen und das Beweisproblem der Betroffenen über Auskunftsansprüche[151] oder Modifikationen der Beweislastverteilung[152] beheben. All diese Ansätze sind im Kern bereits heute umsetzbar. Teilweise bedürfen sie lediglich einer entsprechenden Interpretation bestehender dogmatischer Strukturen (Auskunftsansprüche und Beweislastverteilung), teilweise bewegen sie sich innerhalb des Rahmens, den die Unionsrechtsordnung den Mitgliedstaaten eingeräumt hat (Kollektivverfahren und behördlicher Rechtsschutz).

4. Zwischenfazit

Der Einsatz von ADM-Systemen hält für die bestehende Dogmatik des Nichtdiskriminierungsrechts eine Reihe von Herausforderungen bereit, die bereits heute angegangen werden können. Der Anwendungsbereich des Rechtsgebiets erfasst die meisten der bislang diskutierten Anwendungsfälle. Im Beschäftigungssektor, weil der Einsatz der entsprechenden Systeme keine oder kaum besondere Fragen aufwirft, und im allgemeinen Vertragsrecht, wenn man die entsprechenden dogmatischen Begrifflichkeiten kontextualisiert und technik-responsiv interpretiert. Die Bestimmung der Anforderungen an Diskriminierungstatbestand und Rechtfertigungsprüfung verlangen nach interdisziplinären bzw. interdisziplinärinformierten Antworten. Das von ADM-Systemen ausgehende Transparenzrisiko verschärft das altbekannte Rechtsdurchsetzungsdefizit des Nichtdiskrimi-

---

[148] *Gerards/Xenidis* (Fn. 70), S. 143–144.

[149] Zu dieser Einschätzung *Hacker* (Fn. 20) S. 1167–1183; *Xenidis/Senden* (Fn. 20) S. 175–178; *Gerards/Xenidis* (Fn. 70), S. 144–146; *Grünberger* (Fn. 87). Allgemein zum Problem außerhalb des ADM-Kontextes *Beigang* et al., Möglichkeiten der Rechtsdurchsetzung des Diskriminierungsschutzes bei der Begründung, Durchführung und Beendigung zivilrechtlicher Schuldverhältnisse, 2021, 263–280.

[150] Exemplarisch *Xenidis/Senden* (Fn. 20) S. 178–181; *Gerards/Xenidis* (Fn. 70), S. 145–146. Siehe zum Vorschlag eines behördlichen Rechtsschutzes auch *Hacker* (Fn. 20), S. 1174–1183.

[151] Exemplarisch *Hacker* (Fn. 20), S. 1173–1174.

[152] Dazu etwa *Grünberger* (Fn. 87). Ebenso *Gerards/Xenidis* (Fn. 70), S. 144–145.

nierungsrechts. Ein modifizierter Individualrechtsschutz ist um Instrumente des Kollektivrechtsschutzes und des behördlichen Rechtsschutzes zu ergänzen, um dem Benachteiligungsverbot zu voller praktischer Wirksamkeit zu verhelfen.

IV. Konzeptionelle Perspektiven für morgen

Das Nichtdiskriminierungsrecht verlangt darüber hinaus aber auch nach ganz neuen, konzeptionellen wie dogmatischen Perspektiven, um der besonderen Form und den spezifischen Risiken algorithmischer Diskriminierung gerecht werden zu können. Diesen letzten Teil der Ausgangshypothese will ich anhand zweier Strukturprinzipien[153] des Nichtdiskriminierungsrechts veranschaulichen: Der Reichweite des Rechtsgebiets einerseits (IV.1.) und dem Begriff der (mittelbaren und unmittelbaren) Diskriminierung andererseits (IV.2.).

1. Die Reichweite des Nichtdiskriminierungsrechts: Emergente Diskriminierung und der Katalog der verpönten Kategorien

Das deutsche und europäische Nichtdiskriminierungsrecht ist als *spezielles Gleichbehandlungsrecht*[154] konzipiert. Sein Schutz basiert auf der Anknüpfung einer Entscheidung an einen abschließenden[155] Katalog verpönter Kategorien. Meine These ist, dass diese Konzeption der Eigenart algorithmischer Diskriminierung nicht gerecht wird. Unterstellt, der Online-Finanzdienstleister im Eingangsfall hätte seine automatisierte Kreditvergabeentscheidung nicht an die Kategorien Geschlecht, Alter, Muttersprache und Wohnort geknüpft, sondern an eine Vielzahl von Kriterien, zu denen unter anderem die folgenden[156] gehören: Hundehalter*in, Raucher*in, Leasingwagen, Wohnungseigentümer*in. Der Einsatz des ADM-Systems führt jetzt dazu, dass es im Privatrechtsverkehr zu einer Differenzierung anhand einer Kombination völlig neuartiger und auf den ersten Blick auch ganz unverdächtiger Kriterien kommt. Darin realisiert sich das technische Korrelationsrisiko (dazu oben II.2.c.), also die Art und Weise, wie ADM-Systeme ihre Umwelt anhand von Datenmodellen rekonstruieren: Neue Erkenntnisse werden aus der Kombination einer Vielzahl unterschiedlichster, granularer Daten und

---

[153] Für die unmittelbare und mittelbare Diskriminierung *Ellis/Watson*, EU Anti-Discrimination Law, 2. Auflage 2012, 142–155: „*Key concepts*".

[154] Zur Unterscheidung allgemeines/spezielles Gleichbehandlungsrecht *Mahlmann*, in: Mahlmann/Rudolf, Gleichbehandlungsrecht, 2007, S. 101–102.

[155] Explizit EuGH Urt. v. 11.7.2006 – C-13/05, Slg. 2006, I-06467 Rn. 56 – Chacón Navas; EuGH Urt. v. 17.7.2008 – C-3030/06, Slg. 2008, I-05603 Rn. 46 – Coleman; EuGH Urt. v. 18.12.2014 – C-354/13, ECLI:EU:C:2014:2463 Rn. 34–37 – FOA.

[156] Zu den folgenden Kriterien *Federal Trade Commission*, Data Brokers. A Call for Transparency and Accountability, 2014, Appendix, B3–B6.

ihrer Strukturierung anhand aufgespürter Korrelationen abgeleitet. Die entstehenden personenbezogenen Differenzierungen zielen deshalb häufig auf Unterscheidungskategorien ab, die (wie im Beispiel) weder selbst im Nichtdiskriminierungsrecht verpönt sind, noch mit den im Nichtdiskriminierungsrecht verpönten Kategorien im Zusammenhang stehen.[157] Das Ergebnis sind gänzlich neue Formen von Differenzierungen: *emergente Diskriminierungen*.[158]

a) Problemhintergrund: social salience vs. algorithmic salience

Ist das eine „neue" Form personenbezogener Differenzierung und was unterscheidet sie von „klassischen" Diskriminierungen? Bei der Suche nach einer Antwort auf diese Fragen hilft ein Blick in die jüngere (philosophische) Literatur zur Fundierung und Legitimierung des Nichtdiskriminierungsrechts, insbesondere die einflussreichen Ansätze von *Kasper Lippert-Rasmussen*. Dieser sieht ein zentrales Charakteristikum der ethischen Verwerflichkeit von Diskriminierung darin, dass die Ungleichbehandlung auf eine *socially salient group* bezogen ist.[159] Das ist der Fall „if perceived membership […] is important to the structure of social interactions across a wide range of social contexts."[160] Das bringt das Problem der verpönten Kategorien gut auf den Punkt: Das Nichtdiskriminierungsrecht war und ist eine Reaktion darauf, dass bestimmte soziale Kategorien menschliche, soziale Interaktion in einer Vielzahl von Kontexten strukturieren und dabei Hierarchisierungs- und Exklusionsstrukturen schaffen bzw. aufrechterhalten.[161] Das Konzept der *social salience* kann sowohl in einem objektiven wie auch in einem subjektiven Sinn verstanden werden. Bei *objektivem Verständnis* geht es um die Frage, ob Unterscheidungen anhand einer bestimmten Kategorie aus der Sicht der Öffentlichkeit oder eines verständigen Angehörigen derselben für die Strukturierung sozialer Interaktion von Bedeutung sind.[162] Bei *subjektiver Lesart* steht die Frage im Raum, ob die Kategorie aus der Sicht der betroffenen Gruppe oder ihrer Angehörigen bei der Definition des „personalen Selbst" Relevanz hat.[163]

---

[157] Dies sei zumindest als Prämisse des folgenden Abschnitts unterstellt.

[158] Begriff nach *Mann/Matzner* (Fn. 91), S. 5. Alternative Terminologie bei *Leese*, Security Dialogue 2014, 494 (504): „non-representational categories"; *Mittelstadt*, Philosophy & Technology 2017, 475: „ad hoc groups". Erste Überlegungen zu diesem Phänomen aus rechtlicher Perspektive bei *Zarsky*, (Fn. 6) S. 1405–1411; *Mann/Matzner* (Fn. 91), S. 5–7; *Xenidis/Senden* (Fn. 20), S. 170; *Xenidis* (Fn. 45), S. 751–757; *Wachter*, Berkeley Technology Law Journal 2020, 367 (413–418); *Gerards/Borgesius*, Protected Grounds (Fn. 45).

[159] *Lippert-Rasmussen*, Born Free and Equal? A philosophical inquiry into the nature of discrimination, 2014, insbesondere S. 13–53.

[160] *Lippert-Rasmussen* (Fn. 159), S. 30.

[161] *Grünberger* (Fn. 57), S. 530–534; *Mangold* (Fn. 57), insbesondere S. 397–428. Siehe auch *Balkin*, The Yale Law Journal 1997, 2313.

[162] *Zarsky* (Fn. 4), S. 16.

[163] *Zarsky* (Fn. 4), S. 16.

ADM-Systeme funktionieren anders: Sie interessieren sich bei der Rekonstruktion ihrer gesellschaftlichen Umwelt im Datenmodell nicht für soziale Interaktion bzw. Kommunikation und deren semantische Bedeutung.[164] Sie arbeiten auf rein syntaktischer Ebene und versuchen durch personenbezogene Differenzierungen ein vorgegebenes Ziel zu erreichen, ohne dass es darauf ankäme, dass der Weg dorthin über *socially salient groups* zu erfolgen hat. *Matthias Leese* bringt das gut auf den Punkt, wenn er (allerdings im sicherheitsrechtlichen Kontext) darüber schreibt, wie sich die Logik „gesellschaftlicher Normalität" durch den Einsatz von *Big Data* verändert:

„Starting from the population as the reference point, ‚normality' is no longer defined by social or legal norms, but by the statistical normal distribution of characteristics."[165]

Auf diese Weise werden dynamische und fluide *ad hoc Gruppen*[166] geschaffen, die weder bei objektivem noch bei subjektivem Verständnis mit dem Konzept der *social salience* erfasst werden können. In den Worten von *Tal Zarsky*:

„The individuals are indeed part of a group, but one that is synthetic by nature, of unclear boundaries and structured by an algorithm. It is not a group that the individual feels a strong affinity to, or that the public easily identifies as such."[167]

In Anlehnung an die einflussreiche Terminologie von *Kasper Lippert-Rasmussen* würde ich das Problem deshalb wie folgt umschreiben: Emergente Diskriminierungen (als eine besondere Art statistischer Differenzierungsprozesse) sind nicht weniger rational oder irrational[168] als menschliche Differenzierungen. Sie folgen schlicht anderen Rationalitäten. Es geht nicht mehr länger um die Anknüpfung an *socially salient groups*, auf die das gegenwärtige Nichtdiskriminierungsrecht mit seinem abschließenden Katalog verpönter Kategorien zugeschnitten ist. ADM-basierte Entscheidungen werden vielmehr auf Grundlage von *algorithmically salient groups* getroffen. Die entscheidende Frage für das Recht lautet deshalb: Kann das geltende Nichtdiskriminierungsrecht mit seiner konzeptionellen Anknüpfung an eine abschließende Liste von „single identity ascription[s]"[169] dem Konzept der *algorithmic salience* gerecht werden; und wenn nicht, in welcher Form sollte Abhilfe geschaffen werden? Ich möchte in drei Schritten eine mögliche Antwort darauf kurz skizzieren: Zuerst begründe ich, warum Fälle emergenter Diskriminierung auch aus normativer Sicht problematisch sind (b). In einem zweiten Schritt erkläre ich, warum das geltende Nichtdiskriminierungsrecht die entsprechenden Fälle (noch) nicht erfasst (c). Zuletzt diskutiere ich kurz, welche konzeptionelle(n)

---

[164] Deshalb sieht *Wielsch* (Fn. 64), S. 111 im Datenschutzrecht einen „mittelbare[n] Persönlichkeitsschutz durch die Rechtsverfassung syntaktischer Prozesse".

[165] *Leese* (Fn. 158), S. 501.

[166] Begriff nach *Mittelstadt* (Fn. 158).

[167] *Zarsky* (Fn. 6), S. 1407.

[168] Zu diesem Vorwurf oben II.2.c.

[169] *Grünberger* (Fn. 57), S. 595 und 597.

Ausgestaltung(en) des Nichtdiskriminierungsrechts dem Phänomen gerecht werden können (d).

b) Die normative Relevanz des Problems

Emergente Diskriminierungen durch ADM-Systeme werden in der Literatur überwiegend als normativ bedenklich eingeschätzt: „[They] could still be unfair"[170] bzw. „[They are] unreasonable, counterintuitive, or unjust".[171] Soll eine solche Einschätzung aber nicht nur intuitiven Charakter haben, bedarf sie eines normativen Bewertungsmaßstabs. Für dessen Konkretisierung bietet es sich an, sich an den hier herausgearbeiteten Diskriminierungsrisiken für traditionell geschützte Gruppen zu orientieren: Perpetuierungsrisiko, Generalisierungsrisiko, Korrelationsrisiko und Transparenzrisiko (dazu oben II.2.).[172]

(1.) Identisch ist die Problemlage hinsichtlich des *Generalisierungsrisikos*:[173] Auch im Fall emergenter Diskriminierung werden statistisch valide, aber in Einzelfällen unzutreffende Zuschreibungen von Eigenschaften und Charakteristika vorgenommen, weil Gruppenanknüpfungen imperfekte Generalisierungen auf Grundlage des Induktionsprinzips enthalten.

(2.) Ähnlich liegen die Dinge beim *Transparenzrisiko*:[174] Genauso wie Angehörigen von *socially salient groups* wird es auch den Angehörigen von *algorithmically salient groups* im Prozess schwerfallen, das Vorliegen einer *prima facie Diskriminierung* nachzuweisen. Darüber hinaus verschärft sich das vorgelagerte Problem, Diskriminierung überhaupt als solche zu erkennen: Fluiden, emergenten Gruppierungen fehlt häufig das – den anderen Gruppen aufgrund der geführten Gleichheitskämpfe bewusst gewordene[175] – Wissen, welche Attribute es eigentlich sind, die den Gruppenstatus begründen bzw. die die einzelnen Beteiligten zu Angehörigen der Gruppe machen.[176] Wer aber gar nicht weiß, dass er oder sie unter dem Gesichtspunkt der Zugehörigkeit zu einer „ad hoc Gruppe" diskriminiert wird, kann auch kaum individuellen oder kollektiven, gerichtlichen Rechtsschutz in Anspruch nehmen.

(3.) Komplizierter liegen die Dinge beim *Perpetuierungsrisiko*: Auch hier liegt die Annahme nahe, dass der Einsatz von ADM-Systemen bestehende soziale und ökonomische Gesellschaftsstrukturen aufgreift und in die Zukunft fortschreibt.[177]

---

[170] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 12.

[171] *Wachter* (Fn. 158), S. 416.

[172] Ähnlich *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 12–17.

[173] Ähnlich *Zarsky* (Fn. 6) S. 1409; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 16–17.

[174] Ähnlich *Lees* (Fn. 158), S. 504.

[175] Näher *Mangold* (Fn. 57), S. 185–187.

[176] Dazu *Leese* (Fn. 158), S. 504: „loss of traceability and visibility". Siehe auch *Gandy Jr.* (Fn. 45), S. 40, bei dem allerdings unklar ist, ob emergente oder traditionelle Diskriminierung gemeint ist.

[177] Zu dieser Einschätzung siehe *Zarsky* (Fn. 6), S. 1405–1408: „negative spiral"; *Xenidis* (Fn. 45), S. 751, 753; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 13–15.

Zweifel an der Vergleichbarkeit der Problematik für *socially salient groups* und *algorithmically salient groups* werden jedoch in zweifacher Hinsicht vorgebracht: Erstens sei das Perpetuierungsrisiko dadurch verringert, dass unterschiedliche und voneinander unabhängige ADM-Systeme in verschiedenen sozialen Kontexten zum Einsatz kommen.[178] Anders als bei traditionell geschützten Gruppen, deren Benachteiligung in der Regel eine Vielzahl an Gesellschaftsbereichen durchzieht, würde so verhindert, dass die Benachteiligung einer emergenten Gruppe von einem bestimmten sozialen Kontext auf andere soziale Kontexte übergreift. Es entstehe gerade kein Gruppen-Stigma, sondern lediglich eine zufällige Benachteiligung oder Bevorzugungen konkreter Individuen in bestimmten Situationen. Die damit verbundenen Vor- und Nachteile in spezifischen, sozialen Kontexten können sich dann bei einer Gesamtbetrachtung gegenseitig ausgleichen. Dieser Einwand steht meines Erachtens angesichts aktueller Entwicklungen im Bereich der KI-Forschung auf tönernen Füßen: Wie eine Gruppe von Forscher*innen der Universität Stanford jüngst gezeigt hat, kommt es im Bereich der Entwicklung von ADM-Systemen zunehmend zur Dominanz sogenannter *Foundation Models* (Basismodelle), die bereichsübergreifende Bedeutung erlangen.[179] Entwickler*innen von ADM-Systemen arbeiten danach häufig mit denselben Basismodellen, die dann lediglich punktuell an den jeweiligen sozialen Kontext angepasst werden. Diskriminierende Strukturen können sich auf diesem Wege über eine Vielzahl sozialer Kontexte ausbreiten.[180] Das „Gruppen-Stigma" haftet der „ad hoc"-Gruppierungen dann zwar nicht in den Augen der sonstigen menschlichen Teilnehmer*innen am Rechtsverkehr an, dafür aber nach der Handlungslogik der für die Entscheidung zuständigen ADM-Systeme. Als zweiter Einwand wird geltend gemacht, dass die negativen psychischen Folgen der Benachteiligung für das einzelne Individuum nicht vergleichbar groß seien, wie bei einer Anknüpfung an traditionelle Kategorien.[181] Das ist bei einer subjektiven Lesart des Konzepts der *social salience* verständlich: Wenn keine Identifikation der einzelnen Individuen mit der entsprechenden „ad hoc"-Gruppe stattfindet, wird auch die eigenverantwortliche Definition des „personalen Selbst" weniger beeinträchtigt. Auch diese Erklärung überzeugt als bloße Behauptung ohne empirische Anhaltspunkte nicht. Genauso ließe sich nämlich argumentieren, dass emergente Diskriminierungen für die Betroffenen noch mehr den Anschein von Willkür haben[182] und damit ähnlich große negative, psychische Kosten hervorrufen. Im Ergebnis halte ich deshalb die Perpetuierungsrisiken für traditionelle und emergente Gruppen im Wesentlichen für vergleichbar.

---

178 Zu diesem Argument *Zarsky* (Fn. 6), S. 1408.

179 *Liang* et al., arXiv:2108.07258v2, Edition 18.8.2021.

180 Zum Diskriminierungspotential von *Foundation Models* im Bereich der *large language models* instruktiv *Abid*/*Farooqi*/*Zou*, Nature Machine Intelligence 2021, 461.

181 Zu diesem Argument *Zarsky* (Fn. 6) S. 1407.

182 Vgl. die Nachweise in Fn. 170 und Fn. 171.

(4.) Im Hinblick auf das *Korrelationsrisiko* fällt der Vergleich wieder einfacher: Noch mehr als bei Betroffenheit traditionell geschützter Gruppen wird hier in der Literatur hervorgehoben, dass Differenzierungen nach (augenscheinlich) irrelevanten oder irrationalen Kriterien getroffen werden.[183] Wie bereits betont sollte diese in der Literatur verwendete Terminologie allerdings nicht darüber hinwegtäuschen, dass es nicht um „die eine" Dichotomie von Rationalität und Irrationalität geht, sondern um eine Kollision unterschiedlicher sozialer Rationalitäten (dazu oben II.2.c.). Diese Problemlage wird im Fall emergenter Diskriminierung nur dadurch verdeckt, dass ADM-Systeme auf Kriterien rekurrieren, denen bei der Strukturierung menschlicher, sozialer Interaktion keinerlei Bedeutung zukommt. Das führt aber auch dazu, dass die Differenzierungsprozesse der Öffentlichkeit und den Betroffenen überwiegend verborgen bleiben. In der Folge unterbleibt eine demokratische bzw. gesellschaftliche Kontrolle der Frage, ob die unterschiedlichen Behandlungen durch ADM-Systeme gesellschaftlich erwünscht oder unerwünscht sind.[184] Diese Kontrolle zu leisten ist schon bei der Anknüpfung an traditionell verpönte Kategorien die zentrale Aufgabe des Nichtdiskriminierungsrechts.[185] Im Fall emergenter Diskriminierung gilt das umso mehr. Aus diesen Gründen halte ich auch das *Korrelationsrisiko* für traditionelle und emergente Gruppen im Wesentlichen für vergleichbar.

Legt man den hier gewählten Beurteilungsmaßstab zugrunde, ist die in der Literatur verbreitete Einschätzung, emergente Diskriminierung sei „unfair"[186], „unreasonable, counterintuitive, or unjust"[187] durchaus gerechtfertigt. Ihre normative Bedenklichkeit folgt daraus, dass die von ihr ausgehenden Diskriminierungsrisiken mit denjenigen einer ADM-basierten Diskriminierung traditionell geschützter Gruppen im Wesentlichen vergleichbar sind.

c) Defizite im geltenden Recht

Trotz dieser Vergleichbarkeit der spezifischen Diskriminierungsrisiken werden die entsprechenden Sachverhaltskonstellationen nicht vom Anwendungsbereich des Nichtdiskriminierungsrechts erfasst: Entscheidet ein ADM-System wie im modifizierten Beispiel des Online-Finanzdienstleisters nach emergenten Differenzierungskriterien (Hundehalter*in, Raucher*in, Leasingwagen, Wohnungseigentümer*in), liegt keine rechtlich relevante Diskriminierung vor. Es wird nämlich weder an Kategorien angeknüpft, die selbst im Nichtdiskriminierungsrecht verpönt sind,

---

[183] Zu dieser Einschätzung siehe *Zarsky* (Fn. 6), S. 1408–1411: „arbitrariness-by-algorithm"; *Gerards/Borgesius*, Protected Grounds (Fn. 45). Ähnlich *Xenidis* (Fn. 45), S. 753, die anstatt von Rationalität von Kausalität spricht.

[184] Vgl. dazu auch *Xenidis* (Fn. 45), S. 753.

[185] Dazu *Grünberger* (Fn. 57), v. a. S. 802–804: „Recht auf Rechtfertigung".

[186] Vgl. *Gerards/Borgesius* (Fn. 170).

[187] Vgl. *Wachter* (Fn. 171).

noch an solche, die mit den im Nichtdiskriminierungsrecht verpönten Kategorien in Zusammenhang stehen.[188] Der Fall bewegt sich damit außerhalb des Nichtdiskriminierungsrechts, weil dieses mit seinem abschließenden Katalog verpönter Kategorien auf Differenzierungen nach *socially salient groups* zugeschnitten ist und mit dem Konzept der *algorithmic salience* (bislang) nichts anfangen kann.

d) Konzeptionelle Perspektiven für morgen

Emergente Diskriminierungen verlangen deshalb nach neuen, anders gelagerten normativen Antworten als bisherige Formen personenbezogener Differenzierungen.[189] Man kann diese Antworten einerseits im Nichtdiskriminierungsrecht selbst suchen. Die zentrale Frage lautet dann: Welcher konzeptionelle Ansatz bzw. welches „System" des Nichtdiskriminierungsrechts wird dem Phänomen der emergenten Diskriminierung am ehesten gerecht?[190] Man kann sich andererseits aber auch die (gewissermaßen vorgelagerte) Frage stellen, ob das Nichtdiskriminierungsrecht als Rechtsregime überhaupt das geeignete normativ-regulatorische Instrument ist, um dieser Art personenbezogener Differenzierung zu begegnen.[191] Die Beantwortung der zweiten Frage hängt entscheidend davon ab, ob sich auf erster Stufe ein „adäquates System" des Nichtdiskriminierungsrechts identifizieren lässt.

aa) Konzeptionelle Ansätze und ihre dogmatische Umsetzung

Denkbare konzeptionelle Ansätze zur Ausgestaltung des Rechtsgebiets sind ein offenes System,[192] ein geschlossenes System,[193] sowie ein hybrides System;[194] letzteres als Mischform der beiden erstgenannten. Das *geschlossene System* entspricht der gegenwärtigen Ausgestaltung des deutschen Nichtdiskriminierungsrechts im AGG sowie des europäischen Nichtdiskriminierungsrechts in Form der Richt-

---

[188] Dies sei jedenfalls als Prämisse unterstellt.

[189] Zu dieser Einschätzung siehe auch *Zarsky* (Fn. 6), S. 1405; *Mann/Matzner* (Fn. 91), S. 6. Aus der Perspektive des Sicherheitsrechts *Leese* (Fn. 158), S. 504.

[190] Zu dieser Fragestellung ausführlich *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 4 et passim.

[191] Vgl. dazu auch *Mann/Matzner* (Fn. 91), S. 5, die die Frage aufwerfen, ob in diesen Fällen überhaupt von „Diskriminierung" gesprochen werden sollte. Ähnlich *Martini* (Fn. 90), S. 238–239, der davon ausgeht, dass einfache Hinweispflichten hinsichtlich des Einsatzes von ADM-Systemen ausreichen.

[192] Synonyme Bezeichnungen bei *Fredman* (Fn. 52), S. 118: „*open-textured model*"; *Heringa*, in: Loenen/Rodrigues (Hrsg.), Non-Discrimination Law: Comparative Perspectives, 1999, 25 (27–28): „*open model*"; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 27: „*fully open system*".

[193] Synonyme Bezeichnungen bei *Fredman* (Fn. 52), S. 113: „*exhaustive list*"; *Heringa* (Fn. 192), S. 27–28: „*closed model*"; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 28–29: „*fully closed system*".

[194] Synonyme Bezeichnungen bei *Fredman* (Fn. 52), S. 125: „*non-exhaustive list*"; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 29–34: „*hybrid system*".

linien 2000/43/EG, 2000/78/EG, 2004/113/EG und 2006/54/EG: Der Anwendungsbereich ist auf einen gesetzlich abschließenden Katalog verpönter Kategorien beschränkt, der auch nicht im Wege der richterlichen Rechtsfortbildung erweitert werden kann.[195] Zentraler Akteur ist in diesem System die Legislative, in deren Verantwortungsbereich es liegt, die abschließende Liste verpönter Kategorien zu definieren.[196] Die Judikative hat lediglich in Randbereichen einen eigenen Gestaltungsspielraum, insbesondere bei der Definition der Reichweite einzelner Kategorien in Zweifelsfällen.[197] Entgegengesetzt ist die Lage beim *offenen System*, das den Gerichten die Möglichkeit bietet, jedwede Ungleichbehandlung einer Rechtfertigungsprüfung zu unterziehen und zwar unabhängig davon, ob die Differenzierung anhand gesetzlich vordefinierter, verpönter Kategorien stattfindet oder nicht.[198] Beispiele für ein solches, offenes System wären die Gleichbehandlungssätze des Art. 3 Abs. 1 GG sowie des 14th Amendment to the U.S. Constitution. Zentraler Akteur sind hier die Gerichte, die sowohl für die Bestimmung tatbestandlich relevanter Ungleichbehandlungen wie auch für die entsprechenden Rechtfertigungsanforderungen verantwortlich sind.[199] Gewissermaßen „zwischen den Stühlen" stehen *hybride Systeme*, die den Anwendungsbereich des Nichtdiskriminierungsrechts zwar auf einen gesetzlich vorgegebenen Katalog verpönter Kategorien beschränken, diesen Katalog allerdings nicht abschließend, sondern offen ausgestalten.[200] Die benannten verpönten Kategorien haben in diesem Fall Regelbeispielscharakter.[201] Exemplarisch für solch ein hybrides System stehen Art. 21 Abs. 1 EU-GrCh, Art. 14 EMRK, sowie Art. 15 Abs. 1 der Canadian Charta of Rights and Freedoms. In diesen Fällen kommt sowohl der Legislative wie auch der Judikative eine zentrale Stellung zu: Ersterer durch die Vorgabe der gesetzlich benannten verpönten Kategorien und letzterer durch die Möglichkeit, den Katalog benannter Kategorien unter Bindung an politische Wertentscheidungen zu erweitern.[202]

Die Implementierung eines offenen oder eines hybriden Systems für ADM-basierte Diskriminierung ist *de lege ferenda* Aufgabe der europäischen oder deutschen Gesetzgebung. Aber auch *de lege lata* scheint die dogmatische Umsetzung der entsprechenden konzeptionellen Ansätze zumindest nicht gänzlich ausgeschlossen: Ein denkbarer Ansatzpunkt für die Einführung eines *hybriden Systems*

---

[195] Zum Charakter geschlossener Systeme siehe *Fredman* (Fn. 52), S. 113–118; *Heringa* (Fn. 192), S. 27.

[196] Dazu *Fredman* (Fn. 52), S. 112; *Heringa* (Fn. 192), S. 28; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 38.

[197] Vgl. *Fredman* (Fn. 52), S. 113.

[198] Zum Charakter offener Systeme siehe *Fredman* (Fn. 52), S. 118–125; *Heringa* (Fn. 192), S. 27.

[199] Dazu *Fredman* (Fn. 52), S. 112; *Heringa* (Fn. 192), S. 28; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 41.

[200] Zum Charakter hybrider Systeme siehe *Fredman* (Fn. 52), S. 125–130.

[201] Zum Regelbeispielscharakter siehe *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 31.

[202] Dazu *Fredman* (Fn. 52), S. 112.

wäre die Anwendung von Art. 21 EU-GRC im Horizontalverhältnis.[203] Innerhalb ihres Anwendungsbereichs öffnen die europäischen Nichtdiskriminierungsrichtlinien selbst das Tor des Art. 51 Abs. 1 S. 1 EU-GRC. Außerhalb ihres Anwendungsbereichs könnte ein künftiger AI-Act tauglicher Anknüpfungspunkt für die Eröffnung des Anwendungsbereichs der Grundrechtecharta sein, weil er sachbereichsunabhängig (!) zu einer „diskriminierungsrechtlichen *agency-Situation*"[204] führen würde. Mit der bloßen Eröffnung des Anwendungsbereichs der Charta ist freilich noch kein Wort über die Horizontalwirkung des Art. 21 EU-GRC verloren: Dogmatische Stütze hierfür könnten die Rechtssachen *Egenberger*[205] und *IR*[206] bieten.[207] Ein Ansatzpunkt für die Einführung eines *offenen Systems* könnte hingegen im allgemeinen Europäischen Gleichbehandlungsgrundsatz im Horizontalverhältnis gesucht werden.[208] Dieser hat seine Ausprägung insbesondere in den Rechtssachen *Mangold*[209] und *Kücükdeveci*[210] gefunden. Beide Ansätze in rechtsdogmatisch stimmiger Weise auszubuchstabieren wäre eine wichtige Aufgabe für die Rechtswissenschaft. Entscheidend für die hier untersuchte These ist allein, dass alternative konzeptionelle Lösungsansätze *de lege ferenda* (vielleicht auch schon *de lege lata*) zumindest denkbar sind.

Ob diese Lösungsansätze unterm Strich auch überzeugend sind, hängt primär von den Vorzügen und Nachteilen der einzelnen „Systeme" ab, die einander spiegelbildlich entsprechen. Sie umfassen vier Dimensionen und sind auf abstrakt-genereller Ebene schnell berichtet:[211] (1.) Die erste Dimension betrifft das Verhältnis von Rechtssicherheit und Einzelfallgerechtigkeit:[212] Geschlossene Systeme versprechen ein hohes Maß an Rechtssicherheit, weil klar erkennbar ist, welche Ungleichbehandlungen gesteigerten Rechtfertigungsanforderungen unterliegen.[213] Offene Systeme hingegen bieten flexible Anpassungsmöglichkeiten an die Bedürfnisse einer modernen, sich permanent wandelnden Gesellschaft.[214] Sie können Tatbestands- und Rechtfertigungsvoraussetzungen flexibel an die konkrete, soziale Konfliktlage anpassen und dadurch einzelfallgerechte Entscheidun-

---

[203] Vgl. dazu auch *Xenidis* (Fn. 45), S. 756.

[204] Grundlegend EuGH Urt. v. 13.7.1989 – 5/88, Slg. 1989, 02609, Rn. 19 – Wachauf.

[205] EuGH Urt. v. 17.4.2018 – C-414/16, ECLI:EU:C:2018:257 – Egenberger.

[206] EuGH Urt. v. 11.9.2018 – C-68/17, ECLI:EU:C:2018:696 – IR.

[207] Ob der EuGH allerdings auch bereit wäre, die Horizontalwirkung von Art. 21 GRC außerhalb der Reichweite der in den Nichtdiskriminierungsrichtlinien verpönten Kategorien anzuwenden, ist bislang unklar. Zweifelnd insoweit *Xenidis* (Fn. 45), S. 756.

[208] Vgl. dazu auch *Xenidis* (Fn. 45), S. 756–757.

[209] EuGH Urt. v. 22.11.2005 – C-144/04, Slg. 2005, I-09981 – Mangold.

[210] EuGH Urt. v. 19.1.2010 – C-555/07, Slg. 2010, I-00365 – Kücükdeveci.

[211] Vgl. zu einigen der folgenden Punkte ausführlich *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 37–55.

[212] Grundlegend dazu für den Bereich des Privatrechts *Auer*, Materialisierung, Flexibilisierung, Richterfreiheit, 2005, 46–58: „formale[r] Grundwiderspruch des Privatrechtsdenkens".

[213] Vgl. nur *Grünberger* (Fn. 57), S. 856; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 42.

[214] *Heringa* (Fn. 192), S. 36; *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 55.

gen ermöglichen.[215] (2.) Die zweite Dimension betrifft die Rolle und das Selbstverständnis der Gerichte in der Rechtsordnung und deren Verhältnis zu anderen Gewalten nach dem Gewaltenteilungsgrundsatz:[216] Gegen offene Systeme wird ins Feld geführt, die Gerichte seien nicht legitimiert, originär politische Fragestellungen zu beantworten, worunter dann implizit wohl auch die Definition verpönter Ungleichbehandlungen gefasst wird.[217] Andererseits gilt es zu bedenken, dass es auch originäre Aufgabe der Judikative ist, einzelne Rechtssubjekte vor dem fremdbestimmenden Willen der (in diesem Fall Privatrechts-)Gesellschaft zu schützen.[218] Für geschlossene Systeme wird darüber hinaus ins Feld geführt, der Judikative falle es schwer, einerseits flexible, aber gleichzeitig auch kohärente Standards für die Rechtfertigung von Ungleichbehandlungen zu definieren.[219] Die Legislative sei hingegen schneller im Stande, solche präzisen Standards zu setzen.[220] Dass diese Argumente im Fall politisch hoch brisanter Fragen ein unrealistisch rosiges Bild der gesetzgebenden Gewalt zeichnen, sollte spätestens mit der Debatte um die Einführung einer allgemeinen Gleichbehandlungsrahmenrichtlinie[221] klar geworden sein. Mehr als eine Dekade an Gesetzgebungsprozess sprechen wohl für sich. (3.) Die dritte Dimension zielt auf den „symbolischen Gehalt" des Nichtdiskriminierungsrechts:[222] Offenen Systemen wird ein geringerer „symbolischer Wert" zugeschrieben. Zum einen, weil die verpönten Ungleichbehandlungen nicht durch eine konkrete Benennung verpönter Kategorien explizit gemacht werden. Zum anderen, weil ihre Bestimmung durch die (schwach demokratisch legitimierten) Gerichte stattfindet. Ob beides den „symbolischen Wert" der Definition einer verpönten Ungleichbehandlung tatsächlich beeinträchtigt, erscheint mir allerdings eher zweifelhaft. Man denke etwa nur an die Aufmerksamkeit, die höchstgerichtliche Entscheidungen zu Gleichbehandlungsfragen in der allgemeinen Wahrnehmung und medialen Berichterstattung in jüngerer Zeit erfahren haben.[223] (4.) Die vierte Dimension betrifft praktische Fragestellungen, insbesondere die Frage, ob das Nichtdiskriminierungsrecht im Stande ist, Fälle gesellschaftlicher Ungleichbehandlung in seine eigene Sprache des Rechts „zu übersetzen" und den zugrundeliegenden Konflikt adäquat zu rekonstruieren: Diese Fähigkeit wird geschlossenen

---

[215] Vgl. dazu auch *Thomsen*, Social Theory and Practice 2013, 120 (145).

[216] Grundlegend dazu für den Bereich des Privatrechts *Auer* (Fn. 212), S. 64–93: „institutionelle[r] Grundwiderspruch des Privatrechtsdenkens". Für das Nichtdiskriminierungsrecht *Heringa* (Fn. 192), S. 25, 37.

[217] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 38 und 40–41.

[218] Vgl. *Fredman* (Fn. 52), S. 111.

[219] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 42–43.

[220] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 38 und 42–43.

[221] Vorschlag für eine Richtlinie des Rates zur Anwendung des Grundsatzes der Gleichbehandlung ungeachtet der Religion oder der Weltanschauung, einer Behinderung, des Alters oder der sexuellen Ausrichtung, KOM(2008)426 endg.

[222] Dazu *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 39.

[223] Paradigmatisch etwa BGH Urt. v. 5.5.2021 – VII ZR 78/20, Rn. 22 = NJW 2021, 2514.

Systemen vor allem für die Fälle der Mehrfachdiskriminierung, insbesondere der intersektionellen Diskriminierung abgesprochen.[224]

bb) Hybrides oder offenes System für emergente, ADM-basierte Diskriminierung?

Welche konzeptionelle Ausgestaltung des Nichtdiskriminierungsrechts ist nun vorzugswürdig? Teilweise wird in der Literatur davon ausgegangen, hybride Systeme seien generell überlegen, weil sie die Vorzüge von offenen und geschlossenen Systemen kombinieren und die jeweiligen Nachteile weitestgehend reduzieren würden.[225] Richtigerweise wird man die Frage nach der „generellen Überlegenheit" eines der verschiedenen Systeme auf abstrakt-genereller Ebene nicht beantworten können, weil es stets um eine Abwägung der einzelnen Vorzüge und Nachteile entlang der Achsen der genannten vier Dimensionen geht. Entscheidend ist dabei der konkrete soziale Kontext dieser Abwägung und dieser besteht vorliegend in den spezifischen Besonderheiten emergenter Diskriminierung durch ADM-Systeme: der Anknüpfung an *algorithmically salient groups* anstatt von *socially salient groups*.

Im Folgenden werde ich die Eckpunkte dieser Abwägungsentscheidung kurz umreißen, um zu zeigen, wie die rechtswissenschaftliche Diskussion geführt werden sollte. Im Ausgangspunkt dürfte schnell Einigkeit darüber herzustellen sein, dass das gegenwärtige *geschlossene System* der besonderen Form emergenter Diskriminierung nicht gerecht wird, weil es dieses normativ bedenkliche Phänomen nicht einmal ansatzweise in der Sprache des Nichtdiskriminierungsrechts rekonstruieren kann. *Jannecke Gerards* und *Frederik Borgesius* haben deshalb kürzlich für die Einführung eines *hybriden Systems* plädiert.[226] Die entscheidenden Vorteile gegenüber einem offenen System sehen sie erstens in dem vermeintlich größeren symbolischen Gehalt, zweitens in der effektiven Möglichkeit das Korrelationsrisiko zu adressieren und drittens in der größeren Rechtssicherheit, die ein hybrides System verspricht. Das letzte Wort dürfte damit noch nicht gesprochen sein: Dass offenen Systemen ein geringerer symbolischer Gehalt zukommt als geschlossenen oder hybriden Systemen, ist (wie dargelegt) zumindest zweifelhaft. Das Problem des Korrelationsrisikos, also das Problem der Anknüpfung an (vermeintlich) irrelevante oder irrationale Kriterien (dazu oben II.2.c. und IV.1.b.), kann durch ein offenes System meines Erachtens sogar noch umfassender adressiert werden als durch ein hybrides System. Zuletzt vermag mich das Argument größerer Rechtssicherheit auch nicht recht zu überzeugen. Auf abstrakter Eben mag es zutreffen, dass hybride Systeme ein größeres Maß an Rechtssicherheit garantieren, weil die Regelbeispiele den Gerichten Anhaltspunkte dafür geben,

[224] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 43–44.
[225] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 58–60.
[226] *Gerards/Borgesius*, Protected Grounds (Fn. 45), S. 62–64.

welche neuen Diskriminierungskategorien als verpönt einzuordnen sind. Diesen Gedanken ohne Weiteres auf das Phänomen emergenter Diskriminierungen durch ADM-Systeme zu übertragen, würde jedoch deren spezifische Charakteristika ignorieren: Der bestehende Katalog verpönter Kategorien besteht aus einer Liste an *socially salient groups*. Anhaltspunkte für neue Kategorien vermag er deshalb zwar zu liefern, wenn neuartige, menschliche Differenzierungen ebenfalls auf das Konzept der *social salience* aufbauen. Für *algorithmically salient groups* ist mit Analogien zum bestehenden Katalog hingegen recht wenig gewonnen. Ein Vergleich traditioneller Kategorien wie „Rasse", Alter oder Geschlecht mit emergenten Kategorie-Konglomeraten wie Hundehalter*in, Raucher*in, Leasingwagen, Wohnungseigentümer*in verspricht geringen Mehrwehrt. Weder sind die einzelnen Kriterien vergleichbar, noch kann ein Vergleich der Tatsache Rechnung tragen, dass die emergente Diskriminierung gerade darauf beruhen, dass sich eine Vielzahl (!) vermeintlich unbedenklicher Kategorien überlagert. Der spezifische Mehrwert eines hybriden Systems gegenüber einem offenen System ist es, durch die Regelbeispielstechnik Anhaltspunkte für die Einordnung als bedenkliche oder unbedenkliche Differenzierungsform zu liefern. Dieser Mehrwert scheint mir im Bereich ADM-basierter emergenter Diskriminierung gerade nicht gegeben zu sein.

Auch aus dogmatischer Sicht bedarf die Änderung des konzeptionellen Charakters des Nichtdiskriminierungsrechts weiterer Überlegungen, um Friktionen zu vermeiden. Mit einer bloßen Öffnung des Anwendungsbereichs durch Implementierung eines hybriden oder offenen Systems (für ADM-Systeme) ist das Problem also nicht gelöst. Die dogmatische Ausgestaltung von Diskriminierungstatbestand und Rechtfertigungsmöglichkeiten ist nämlich auf das gegenwärtige, geschlossene System zugeschnitten und müsste angepasst werden. Hierzu nur einige Andeutungen, um das Problem knapp zu veranschaulichen: Die konzeptionelle Dichotomie zwischen unmittelbarer und mittelbarer Diskriminierung scheint mir kaum mehr haltbar zu sein, wenn (sowohl beim hybriden wie auch beim offenen System) grundsätzlich jedes beliebige Differenzierungskriterium eine Diskriminierung potentiell begründen soll (dazu unten IV.2.b.). Darüber wäre es erforderlich, das Verhältnis zwischen einem (engen oder weiten) Diskriminierungsbegriff und (engen oder weiten) Rechtfertigungsmöglichkeiten neu auszutarieren. Zuletzt muss auch die (Nicht-)Anwendbarkeit bestimmter dogmatischer Figuren auf Fälle emergenter Diskriminierungen erwogen werden. Zu denken ist hier insbesondere an Positivmaßnahmen nach § 5 AGG. Diese sind Ausdruck eines materialen Gleichheitsverständnisses und dienen der Bekämpfung von lang andauernden, sozialen, strukturellen Nachteilen.[227] Sie vor diesem teleologischen Hintergrund auch auf emergente Diskriminierungen anzuwenden, liegt zumindest nicht evident auf der Hand.

---

[227] *Grünberger* (Fn. 57), S. 710–713.

e) Fazit

Die Frage, wie auf das neuartige Phänomen emergenter Diskriminierung durch ADM-Systeme normativ-regulatorisch reagiert werden sollte, ist zentral für die Zukunft des Nichtdiskriminierungsrechts. Diese Form der Diskriminierung knüpft an *algorithmically salient groups* und gerade nicht mehr an *socially salient groups* an. Sucht man eine Antwort auf die Frage innerhalb des Nichtdiskriminierungsrechts, gilt es zwischen einem hybriden und einem offenen System abzuwägen. Die (vermeintlichen) Vorteile eines hybriden Systems sind im konkreten Fall emergenter, ADM-basierter Diskriminierung zweifelhaft. Es ist deshalb erwägenswert, über einen *allgemeinen Gleichbehandlungsgrundsatz* für ADM-Systeme nachzudenken (offenes System). Die Nachteile einer Ausdehnung der Reichweite des Diskriminierungsverbots im Privatrechtsverkehr können nämlich nicht nur über eine Restriktion der verpönten Kategorien ausgeglichen werden (geschlossenes oder hybrides System). Die Reichweite des Benachteiligungsverbots könnte ebenso über eine Begrenzung eines allgemeinen Gleichbehandlungsgrundsatzes auf bestimmte soziale Kontexte (Miete, Versicherung, Kreditwirtschaft etc.) feinjustiert werden. Vorteil dieser Lösung wäre, dass dann nicht mehr die (für emergente Diskriminierung untauglichen) Kategorien der *socially salient groups* über die Anwendbarkeit des Diskriminierungsverbotes entscheiden, sondern die Frage, ob bestimmte soziale Kontexte eine erhebliche Relevanz für die allgemeine Lebensführung der betroffenen Personen haben. Als dritte Option bleibt freilich immer auch die Möglichkeit, die Antwort auf die Ausgangsfrage nicht im Nichtdiskriminierungsrecht zu suchen, sondern diese Aufgabe an alternative Formen der KI-Regulierung zu delegieren.

2. Der Begriff der Diskriminierung: Mittelbare und unmittelbare Diskriminierung auf dem Prüfstand

Der Diskriminierungsbegriff ist das zweite Strukturprinzip, um meine Ausgangshypothese zu veranschaulichen. Das europäische und deutsche Nichtdiskriminierungsrecht kennen insgesamt fünf Tatbestände der Diskriminierung, wobei der mittelbaren und der unmittelbaren Diskriminierung die mit Abstand größte Bedeutung zukommt.[228] Eine *unmittelbare Diskriminierung* liegt vor, wenn „eine Person aufgrund [einer im Nichtdiskriminierungsrecht verpönten Kategorie] in einer vergleichbaren Situation eine weniger günstige Behandlung als eine andere Person erfährt, erfahren hat oder erfahren würde".[229] Von einer *mittelbaren Dis-*

[228] Daneben kennt das Nichtdiskriminierungsrecht noch die Kategorien der Belästigung (vgl. § 3 Abs. 3 AGG), der sexuellen Belästigung (vgl. § 3 Abs. 4 AGG) und der Anweisung zur Benachteiligung (vgl. § 3 Abs. 5 AGG).

[229] Vgl. Art. 2 Abs. 2 lit. a RL 2000/43/EG; Art. 2 Abs. 2 RL 2002/73/EG; Art. 2 Abs. 2 lit. a RL 2000/78/EG; Art. 2 lit. a RL 2004/113/EG; § 3 Abs. 1 AGG.

kriminierung wird hingegen gesprochen, wenn „dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen wegen [einer im Nichtdiskriminierungsrecht verpönten Kategorie], in besonderer Weise benachteiligen können, es sei denn, die betreffenden Vorschriften, Kriterien oder Verfahren sind durch ein rechtmäßiges Ziel sachlich gerechtfertigt, und die Mittel sind zur Erreichung dieses Ziels angemessen und erforderlich".[230] Rechtsdogmatische und rechtspraktische Bedeutung erfährt die Unterscheidung dieser beiden Diskriminierungskategorien vor allem bei der Frage nach der Art und den Anforderungen an die Beweisführung[231] sowie bei der Intensität der Rechtfertigungsprüfung[232]. Meine diesbezügliche These lautet, dass die spezifische Funktionsweise von ADM-Systemen dazu führt, dass die Grenze zwischen den rechtsdogmatischen Diskriminierungskategorien der unmittelbaren und mittelbaren Diskriminierung zu verschwimmen droht. Die strikte tatbestandliche Dichotomie des gegenwärtigen Nichtdiskriminierungsrechts ist auf konzeptioneller Ebene untauglich, um der spezifischen Eigenart ADM-basierter Diskriminierung gerecht zu werden.[233] Zur Verdeutlichung dieser These gehe ich in drei Schritten vor: Zuerst nehme ich die bisherigen Einschätzungen der Literatur zur Bestimmung der Diskriminierungskategorie in den Blick (a). Sodann beleuchte ich den konzeptionellen wie auch sozio-technischen Hintergrund des Problems (b). Zuletzt diskutiere ich, ob eine konzeptionelle Neuausrichtung des Diskriminierungstatbestands Abhilfe schaffen kann (c).

a) Falsche Fronten …

Die Literatur geht bislang ganz einhellig davon aus, dass Ungleichbehandlungen durch ADM-Systeme in der Regel die Rechtsform der mittelbaren Diskriminierung annehmen werden.[234] Diese Einschätzung gründet auf einem für ADM-Systeme charakteristischen Phänomen: Bei den Input-Variablen des algorithmischen Datenmodells wird es sich selten um die verpönten Kategorien selbst handeln, sondern um (vermeintlich) neutrale Kriterien, die einzeln, vor allem aber in ihrer Summe, mit verpönten Kategorien korrelieren (können).[235] Für die unmittel-

---

[230] Vgl. Art. 2 Abs. 2 lit. b RL 2000/43/EG; Art. 2 Abs. 2 RL 2002/73/EG; Art. 2 Abs. 2 lit. b RL 2000/78/EG; Art. 2 lit. b RL 2004/113/EG; § 3 Abs. 2 AGG.

[231] Dazu (unter anderem auch aus rechtsvergleichender Perspektive) *Santagata*, ZEASR 2020, 253.

[232] Siehe hierzu *Fredman* (Fn. 52), S. 190–202.

[233] Andeutungsweise auch *Gerards/Xenidis* (Fn. 70), S. 67 und 143–144.

[234] Siehe zu dieser Einschätzung aus Sicht des europäischen Nichtdiskriminierungsrechts vor allem *Hacker* (Fn. 20), S. 1153; *Xenidis/Senden* (Fn. 20) S. 171–175; *Borgesius*, The International Journal of Human Rights 2020, 1572 (1576–1578); *Gerards/Xenidis* (Fn. 70), S. 72–73. Aus der Perspektive des deutschen Nichtdiskriminierungsrechts *Dzida/Groh* (Fn. 101), S. 1919–1920; *Freyler* (Fn. 101), S. 287–288; *Kullmann* (Fn. 101), S. 242–243; *von Ungern-Sternberg* (Fn. 139), 14–15. Für das U.S. Nichtdiskriminierungsrecht paradigmatisch *Barocas/Selbst* (Fn. 20), S. 694–712.

[235] *Gerards/Xenidis* (Fn. 70), S. 63.

bare Diskriminierung bleibe deshalb beim Einsatz von ADM-Systeme nur wenig Raum.[236] Diese Einschätzung klingt zunächst plausibel. Erste Zweifel stellen sich aber ein, wenn man bedenkt, dass *Gabriele Britz* vor etwas mehr als einer Dekade für statistische Entscheidungsverfahren geradezu eine „Renaissance unmittelbarer Diskriminierung" prognostiziert hat.[237]

b) Sozio-technische und konzeptionelle Hintergründe des Problems

Erklären lässt sich die Diskrepanz in der Einschätzung mit der spezifischen soziotechnischen Funktionsweise von ADM-Systemen: Diese versuchen, bestimmte Zielvariablen dadurch zu optimieren, dass sie enorme Datensätze – die aus einer Vielzahl unterschiedlicher, granularer Daten bestehen – anhand aufgespürter Korrelationen strukturieren. Gerade darauf beruhen auch zwei der charakteristischen Diskriminierungsformen durch ADM-Systeme: *emergente Diskriminierung* einerseits und *correlation discrimination* andererseits.

aa) Abgrenzungsprobleme bei correlation discrimination

Im Unterschied zu „klassischen" Prozessen statistischer Differenzierung stützen ADM-Systeme ihre Entscheidung nicht mehr auf einzelne oder wenige Kriterien (die häufig selbst verpönt sind), sondern auf eine fast unüberschaubare Vielzahl unterschiedlicher Datenpunkte.[238] Dieser Unterschied erklärt zum einen, warum *Gabriele Britz* die unmittelbare, die gegenwärtige Literatur dagegen die mittelbare Diskriminierung als die zentrale Diskriminierungskategorie statistischer Differenzierungsprozesse identifiziert. Der Vergleich von „klassischer" und ADM-basierter statistischer Diskriminierung offenbart aber auch zwei Gründe, weshalb die Grenzziehung zwischen unmittelbarer und mittelbarer Diskriminierung als solche im Fall der *correlation discrimination* erheblich an Überzeugungskraft einbüßt: (1.) Erstens ist in der Rechtsprechung des EuGH[239] und des BAG[240] seit langem anerkannt, dass die Anknüpfung an (vermeintlich) neutrale Kriterien, die selbst keine verpönte Kategorie sind, eine unmittelbare Diskriminierung begründen kann, wenn diese Kriterien mit einer verpönten Kategorie „in einem untrennbaren Zusammenhang"[241] stehen. In diesem Fall wird häufig von einer verdeckten Diskriminierung gesprochen.[242] Wann dies genau der

---

[236] Zu dieser Einschätzung *Hacker* (Fn. 20), S. 1151–1154.
[237] *Britz* (Fn. 45), S. 58–60.
[238] Ähnlich *Hacker* (Fn. 20), S. 1148–1149; *Orwat* (Fn. 8), S. 31–33.
[239] Grundlegend EuGH Urt. v. 12.10.2010 – C-499/08, Slg. 2010, I-09343 – Andersen.
[240] Grundlegend BAG Urt. v. 7.6.2011 – 1 AZR 34/10 = NZA 2011, 1370.
[241] Zu dieser Formulierung EuGH Urt. v. 12.10.2010 – C-499/08, Slg. 2010, I-09343, Rn. 23 – Andersen.
[242] Vgl. MüKo-BGB/*Thüsing*, 9. Auflage, § 3 AGG Rn. 15; BeckOGK-AGG/*Baumgärtner* (Fn. 39), § 3 Rn. 52–54.

Fall ist, ist schon in den „klassischen" Fällen statistischer Diskriminierung durch Menschen nicht leicht zu bestimmen, liegt aber wohl noch im Bereich des Machbaren: Werden statistische Voraussagen auf der Grundlage einzelner Stellvertretervariablen getroffen, kann jeweils noch mit Hilfe intuitiver Kausalitätserwägungen beurteilt werden, ob die „überproportionale Belastung" zufällig oder gezielt erfolgt – ob die Stellvertretervariablen also „nahe genug dran" sind an der verpönten Kategorie, um als eine unmittelbare Anknüpfung gelten zu können. Wenn zum Beispiel an eine Schwangerschaft[243] oder den familienrechtlichen Status als eingetragene Lebenspartner*innen[244] angeknüpft wird, ist vergleichsweise leicht intuitiv nachvollziehbar (und deshalb unbestritten), dass die Kategorien Geschlecht und sexuelle Orientierung lediglich verdeckt werden. Im Kontext ADM-basierter Entscheidungen ist das anders: Weil hier gerade an die Kumulation einer unüberschaubaren Vielzahl granularer Daten angeknüpft wird, versagen intuitive Kausalitätserwägungen regelmäßig.[245] Es lässt sich nicht mehr sagen, ob die „Benachteiligung in besonderer Weise" lediglich zufällige Auswirkung einer neutralen Anknüpfung ist (mittelbare Diskriminierung) oder ob der Score des ADM-Systems als Stellvertretermerkmal das verpönte Hauptmerkmal lediglich ersetzt (unmittelbare Diskriminierung).[246] (2.) Besonders relevant wird diese Problemlage bei sog. „sozial konstruierten" verpönten Kategorien.[247] Das sind solche, deren Gehalt sich erst aus einer sinnverstehenden Kombination verschiedener Umstände ergibt. Genau diese Umstände können aber auch solche sein, an die das ADM-System anknüpft, ohne die semantische Bedeutung dieses Vorgangs erfassen zu können: Wenn im Amazon-Eingangsfall der Recruitment-Algorithmus exemplarisch anhand der Parameter ‚Sprache' und ‚äußeres Erscheinungsbild' unterscheiden würde, wären das Kriterien, die zwar einerseits (vermeintlich) neutral sind, die aber andererseits wiederum erst die soziale Kategorie „Rasse" erzeugen (können). Die zentrale Frage ist dann, ab welcher quantitativen und qualitativen Schwelle die Grenze zur unmittelbaren Diskriminierung überschritten ist.[248] Diese Frage zu beantworten mag bei der Anknüpfung an eine oder einzelne Stellvertretervariablen noch möglich sein. Im Fall ADM-basierter Entscheidungen mit ihrer Anknüpfung an eine Vielzahl unterschiedlicher granularer Daten wird die Suche nach einer Antwort zu einer Aufgabe, der wohl nur *Ronald Dworkins* Richter Herkules[249] gewachsen sein dürfte.

---

[243] Vgl. dazu § 3 Abs. 1 S. 2 AGG.

[244] Zu diesem Beispiel *Grünberger* (Fn. 57), S. 643–645 und 655–656.

[245] Siehe auch *Prince/Schwarcz* (Fn. 71), S. 1263–1264.

[246] Allgemein zu diesem Problem *Grünberger* (Fn. 57), S. 655–656; *Britz* (Fn. 45), S. 55–58.

[247] Vgl. dazu statt vieler und m. w. N. *Grünberger* (Fn. 57), S. 562–564 („Rasse") und 585–587 (Behinderung).

[248] Diese Frage werfen auch *Gerards/Xenidis* (Fn. 70), S. 63–64 auf.

[249] *Dworkin*, Law's Empire, 1986.

bb) Abgrenzungsprobleme bei emergenter Diskriminierung

Besonders deutlich wird die begrenzte Leistungsfähigkeit der Dichotomie von unmittelbarer und mittelbarer Diskriminierung im Fall *emergenter Diskriminierung*. Das liegt daran, dass die Unterscheidung auf konzeptioneller Ebene auf einer Differenzierung zwischen Individualbezug und Gruppenbezug aufbaut und diese Grenzziehung im Fall emergenter Diskriminierungsformen durch ADM-Systeme verschwimmt. Die rechtsdogmatische Kategorie der unmittelbaren Diskriminierung wird konzeptionell zumeist auf das Prinzip zurückgeführt, gleiche Sachverhalte bzw. Individuen gleich zu behandeln. Das Ziel des Verbots unmittelbarer Diskriminierung wird deshalb vielfach in der Verwirklichung von *individual fairness* gesehen.[250] Die rechtsdogmatische Kategorie der mittelbaren Diskriminierung soll hingegen auf dem Prinzip der Gleichbehandlung verschiedener (sozialer) Gruppen beruhen und der Verwirklichung von *group fairness* dienen.[251] In den Fällen emergenter Diskriminierung ist die Dichotomie dieser Fairnesskonzeptionen allerdings problematisch: Extrem dynamische, hoch granulare Kategorisierungen führen dazu, dass die Verdichtung zu (neuen) sozialen Gruppen fragwürdig wird. Das zeigt das modifizierte Kreditbeispiel: Wird bei der Anknüpfung an die Kriterien Hundehalter*in, Raucher*in, Leasingwagen, Wohnungseigentümer*in eine neue soziale Gruppierung geschaffen oder lediglich ein Individuum als eine Kumulation aus einer unüberschaubaren Anzahl einzelner Kategorisierungen rekonstruiert? Dass die Frage schwer zu beantworten ist, muss nicht bedeuten, dass die Suche nach einer zutreffenden Antwort intellektuelle Höchstleistungen erfordert. Es kann auch einfach ein Hinweis darauf sein, dass die der Frage zugrunde liegende Prämisse (Gruppierungen und Individuen ließen sich im Zeitalter von Big Data noch sinnvoll unterscheiden) falsch gesetzt ist. Hinsichtlich der gegenwärtig geführten Paralleldebatte um *personalized law*[252] bemerken *Omri Ben-Shahar* und *Ariel Porat* zutreffend:

„People are never put into group „bins" under personalized law; each person is a bin of one."[253]

Die hochgradige Granularisierung algorithmischer Differenzierungsprozesse führt also dazu, dass die traditionelle Verknüpfung von mittelbarer Diskriminierung mit Gruppen und unmittelbarer Diskriminierung mit Individuen ihre Aussagekraft verliert. Wenn aber die konzeptionelle Dichotomie aus (sozialer) Gruppe und Individuum keine Rolle mehr spielt, verschwimmt notwendigerweise auch die Grenze

---

[250] Exemplarisch *Hacker* (Fn. 20), S. 1175.

[251] Exemplarisch *Hacker* (Fn. 20), S. 1175.

[252] Siehe zu dieser Debatte die Beiträge des Symposiums *Personalized Law* in The University of Chicago Law Review Vol. 86 No. 2 (2019), sowie *Ben-Shahar*/*Porat*, Personalized Law. Different Rules for Different People, 2021 und *Busch*/*De Franceschi* (Hrsg.), Algorithmic Regulation and Personalized Law, 2021.

[253] *Ben-Shahar*/*Porat* (Fn. 252), S. 163.

zwischen den rechtsdogmatischen Kategorien, die auf dieser Dichotomie aufbauen. Für den Fall emergenter Diskriminierung erweist sich die Unterscheidung von unmittelbarer und mittelbarer Diskriminierung deshalb als ungeeignet.

c) … und konzeptionelle Auswege

Der Einsatz von ADM-Systemen führt also dazu, dass die Grenze zwischen unmittelbarer und mittelbarer Diskriminierung zunehmend verschwimmt. Wie könnte vor diesem Hintergrund eine adäquate Ausgestaltung des Diskriminierungstatbestands für die digitale Gesellschaft aussehen? Eine mögliche Antwort auf dieses Problem wäre es, die zweiwertige Struktur des gegenwärtigen Diskriminierungstatbestands zugunsten eines ‚Einheitsmodells' aufzugeben. Interessanterweise könnte ein solcher Ansatz auf ein historisches Vorbild zurückblicken: Im Jahr 1999 hat der Supreme Court of Canada in der richtungsweisenden Grundsatzentscheidung *B.C. Firefighters*[254] die zweiwertige Unterscheidung zwischen *direct discrimination* und *adverse effect* zugunsten einer einheitlichen Diskriminierungsprüfung aufgegeben. Die Begründung klingt vertraut:

„The distinction between a standard that is discriminatory on its face and a neutral standard that is discriminatory in its effect is difficult to justify, simply because there are few cases that can be so neatly characterized."[255]

Wie ein solches ‚Einheitsmodell' ausgestaltet werden kann, möchte ich kurz anhand des Rechtfertigungsmaßstabs andeuten: Die unterschiedlichen Anforderungen der Rechtfertigung für unmittelbare und mittelbare Diskriminierung werden dabei aufgegeben und von einem einheitlichen Rechtfertigungsmaßstab abgelöst. Damit soll freilich nicht gesagt werden, dass algorithmische Diskriminierungen stets denselben Rechtfertigungsanforderungen unterfallen. Aufgegeben wird allein die Bildung von Maßstäben anhand der rechtsdogmatischen Figuren der mittelbaren und unmittelbaren Diskriminierung. Ins Zentrum rückt stattdessen eine Differenzierung nach den konkreten, faktischen Erscheinungsformen algorithmischer Diskriminierung: Die Anforderungen an die Rechtfertigungsprüfung unterscheiden sich dann je nach Ursache[256] (dazu II.1.) oder konkreter Erscheinungsform[257] der Diskriminierung. Ein solcher Ansatz kann den soziotechnischen Hintergründen und den sozialen Wirkungen ADM-basierter Diskriminierung besser Rechnung tragen, als eine Orientierung an rechtsdogmatischen Figuren, die die zugrundeliegenden faktischen Phänomene kaum mehr adäquat abzubilden vermögen.

---

[254] British Columbia v. BCGSEU, [1999] 3 S.C.R. 3.

[255] British Columbia v. BCGSEU (Fn. 254), S. 18.

[256] Dazu gehören die vier Kategorien *biased training data*, *unequal base rates*, *bias in modeling* und *bias in usage*.

[257] Exemplarisch genannt seien die *emergente Diskriminierung*, die *proxy discrimination* und die *incidental discrimination*.

d) Fazit

Die spezifische, sozio-technische Funktionsweise von ADM-Systemen führt zu einer Verschleifung der dogmatischen Kategorien der mittelbaren und unmittelbaren Diskriminierung. Differenzierende Anforderungen insbesondere an Rechtfertigungsprüfung und die Beweisführung, die nach diesen rechtsdogmatischen Kategorien unterscheiden, überzeugen deshalb im Fall algorithmischer Diskriminierung wenig. Einen konzeptionellen Ausweg könnte ein ‚Einheitsmodell' bieten. Dies auszubuchstabieren ist Aufgabe der Rechtswissenschaft.

V. Fazit

Die Zukunftstechnologie „Künstliche Intelligenz" ist in aller Munde: Wirtschaft, Zivilgesellschaft und die an der Gesetzgebung beteiligten Akteure diskutieren über die Versprechungen algorithmischer Entscheidungssysteme, aber auch über adäquate Regulierungsformen. Die Verheißungen solcher Systeme sind groß, sollen sie doch bei personenbezogenen Entscheidungen Effizienz und Fairness gleichermaßen fördern. Dennoch besteht mittlerweile Einigkeit, dass von ihnen zugleich ein erhebliches Diskriminierungspotential ausgeht. Das Spezifikum dieser Diskriminierung ist ihre Datenbezogenheit: sie kann dem algorithmischen Datenmodell entspringen (*biased training data*), den abgebildeten gesellschaftlichen Realitäten entstammen (*unequal base rates*), auf Eingriffen in das Datenmodell beruhen (*bias in modeling*) oder aber das Resultat einer fehlerhaften Anwendung des Systems sein (*bias in usage*). Kennzeichnend für diese algorithmische Diskriminierung sind vier normativ relevante Diskriminierungsrisiken: Perpetuierungsrisiko, Generalisierungsrisiko, Korrelationsrisiko und Transparenzrisiko. Diese Risiken zu adressieren ist zuvörderst Aufgabe des Nichtdiskriminierungsrechts.

Der Beitrag entwickelt die These, dass die spezifische Form der Diskriminierung durch ADM-Systeme die tragenden Strukturen und Konzepte dieses Rechtsgebiets nachhaltig herausfordert. Einige dieser Herausforderungen können bereits heute mit dem bestehenden dogmatischen Instrumentarium angegangen werden: Der Anwendungsbereich bedarf dafür einer technik-responsiven Interpretation. Diskriminierungstatbestand, Rechtfertigungsmöglichkeiten und Rechtsfolgen verlangen nach interdisziplinär-informierten Antworten unter besonderer Berücksichtigung der verschiedenen Ursachen algorithmischer Diskriminierung. Das Transparenzrisiko verschärft das altbekannte Rechtsdurchsetzungsdefizit des Nichtdiskriminierungsrechts. Der Individualrechtsschutz bedarf deshalb der Modifikation im Hinblick auf Beweisverteilung und/oder Auskunftsansprüche, sowie einer Ergänzung um Instrumente des Kollektivrechtsschutzes und des behördlichen Rechtsschutzes. Nur so kann dem Benachteiligungsverbot auch zu

voller praktischer Wirksamkeit verholfen werden. Mittel- bis langfristig bedarf das Nichtdiskriminierungsrecht aber ganz neuer konzeptioneller und dogmatischer Perspektiven, um der besonderen Form und den spezifischen Risiken algorithmischer Diskriminierung gerecht werden zu können. Die dafür adäquaten Begriffe und Konzepte zu entwickeln, ist Aufgabe der Rechtswissenschaft.[258] Sie muss sich dieser Aufgabe allerdings nicht allein stellen und sollte dies auch nicht tun. Nur eine sozialwissenschaftlich, ethisch und technisch informierte – also responsive[259] – Rechtswissenschaft kann das Phänomen algorithmischer Diskriminierung adäquat in der Sprache des Nichtdiskriminierungsrechts rekonstruieren.

Gleichzeitig möchte ich vor einer überzogenen Erwartungshaltung an das Nichtdiskriminierungsrecht warnen: Der Schlüssel für eine diskriminierungsfreie ‚digitale Gesellschaft' liegt vermutlich nicht im Nichtdiskriminierungsrecht allein, sondern in seiner Einbettung in einen Regulierungsrahmen, der als Gesamtkonzept das Phänomen algorithmischer Diskriminierung bewältigen kann. Dieser sollte auf drei Säulen aufbauen: dem Nichtdiskriminierungsrecht, dem Datenschutzrecht und, drittens, einem künftigen AI-Act, der sich der Diskriminierungsrisiken von ADM-Systemen bewusst ist. Aufgabe der Rechtswissenschaft ist es, das Zusammenspiel von Prozessregulierung (Datenschutzrecht und AI-Act) und Ergebnisregulierung (Nichtdiskriminierungsrecht und AI-Act) zu gestalten, um mit diesem regulatorischen Mix die zentralen Probleme algorithmischer Diskriminierung effektiv adressieren zu können. Gelingt dies, stehen die Chancen nicht schlecht, dass der Einsatz algorithmischer Entscheidungssysteme nicht zu einer Verstärkung diskriminierender Strukturen beitragen wird, sondern zu deren Bekämpfung.

---

[258] Dazu im Kontext der Regulierung von Online-Plattformen *Podszun*, Empfiehlt sich eine stärkere Regulierung von Online-Plattformen und anderen Digitalunternehmen? – Gutachten F zum 73. Juristentag, 2020, F20.

[259] Dazu *Grünberger* (Fn. 107) sowie die weiteren Nachweise in Fn. 107.

Vertrauenswürdige Verwendung von Künstlicher Intelligenz in Deutschland und Europa

Frauke Rostalski

I. Einleitung

Die digitale Transformation der Gesellschaft schreitet voran. In den Fokus rückt dabei insbesondere die Technologie der Künstlichen Intelligenz (KI), die den Lebensalltag von Verbraucherinnen und Verbrauchern mehr und mehr durchdringt. Künstliche Intelligenz versteht sich dabei als Technologie, die das Ziel verfolgt, intelligentes Verhalten, zum Beispiel Wahrnehmen, Schlussfolgern und Entscheiden, zu automatisieren. Hinter diesem Schlagwort verbergen sich – je nach Vorverständnis – ganz unterschiedliche technische Systeme. Oftmals sind Techniken des Maschinellen Lernens (ML) gemeint.[1] Anwendungen, die auf dieser Technik basieren, lernen, auf Basis sog. Trainingsdaten Muster und Regeln zu identifizieren, sie zu verallgemeinern und auf neue Sachverhalte anzuwenden.[2] Aber auch „‚einfache' Regelsysteme, die auf durch Experten ‚manuell' festgelegten Algorithmen (= Regeln) beruhen",[3] werden zu Recht einbezogen.[4] KI-Technologie begegnet der Verbraucherin oder dem Verbraucher nicht nur in ihrem oder seinem eigenen Zuhause (z. B.: Steuerung von Beleuchtung und Wärmezufuhr, Sprachassistenten) oder in der privaten Kommunikation mit anderen, zum Beispiel in sozialen Medien. Darüber hinaus kommen Systeme der Künstlichen Intelligenz in einer Vielzahl weiterer Lebensbereiche zum Einsatz, die Verbraucherinnen und Verbraucher nahezu täglich betreffen: Bei der Inanspruchnahme von Finanzdienstleistungen, im Versicherungswesen oder auf dem Arbeitsmarkt – etwa in Gestalt von KI-Systemen, die Auswahlentscheidungen in Bewerbungsprozessen unterstützen oder eine Risikobewertung der oder des Versicherten übernehmen. Die Künstliche Intelligenz ist vor diesem Hintergrund längst im Leben der Verbraucherinnen und Verbraucher angekommen. Dabei steht die Entwick-

---

[1] Siehe *Hacker*, NJW 2020, 2142 (2142).

[2] *Leupold/Wiesner*, in: Leupold/Wiebe/Glossner, Münchener Anwaltshandbuch IT-Recht, 4. Aufl. 2021, Teil 9.6.4. Rn. 2 m. w. N.

[3] *Datenethikkommission (DEK)*, Gutachten der Datenethikkommission der Bundesregierung 2019, S. 34 (abrufbar unter: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_DE.pdf?__blob=publicationFile&v=5, letzter Zugriff 14.12.2021).

[4] Vgl. zum Begriffsverständnis ausführlich *Rostalski/Weiss*, ZfDR 2021, 329 (331 f.).

lung in vielen Bereichen erst am Anfang. Mit alledem gehen zahlreiche Vorteile einher: KI-Systeme sind geeignet, den Alltag erheblich zu erleichtern. Nicht zu vergessen ist der Zugewinn an Sicherheit, der zum Beispiel durch den Einsatz von KI im Bereich der Mobilität erreicht werden kann. In der Medizin kann KI dazu beitragen, die Behandlung von Menschen erheblich zu verbessern, indem etwa neue Behandlungsmethoden oder Medikamente unter Zuhilfenahme der Technologie entwickelt werden.

Gleichwohl gehen mit KI-Anwendungen im Leben von Verbraucherinnen und Verbrauchern auch Risiken einher. Der Blick fällt dabei zum Beispiel auf den Bereich des Datenschutzes: Da KI-Systeme häufig aus Daten „lernen" und die Menge und Qualität an verfügbaren Daten nicht selten über die Qualität der Anwendungen entscheidet, drohen in besonderem Maße Risiken für die informationelle Selbstbestimmung der oder des Einzelnen, der ggf. der Verwendung seiner Daten nicht zustimmt bzw. nicht einmal Kenntnis davon erlangt. Künstliche Intelligenz ist indessen aus rechtlicher Sicht kein bloßes Datenschutzthema. Aufgrund der ML-Systemen inhärenten Eigenschaft des „Selbstlernens", das die Technologie gar zu einer „Black Box"[5] machen kann, treten oftmals Anforderungen an die Transparenz und Nachvollziehbarkeit der Ergebnisgenerierung durch die jeweilige KI-Anwendung in den Vordergrund. Und nicht erst seit der US-amerikanischen KI-Anwendung COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), die Richterinnen und Richter bei der Entscheidung über die Anordnung von Untersuchungshaft, bei der Strafzumessung sowie bei der Entscheidung über eine Strafaussetzung zur Bewährung unterstützt, ist klar, dass Fragen der Diskriminierung im Kontext von Künstlicher Intelligenz eine hervorgehobene Rolle spielen: Je nachdem, welche Daten verwendet und wie diese bewertet werden, kann die Nutzung des KI-Systems diskriminierende Folgen haben, die von Rechts wegen nicht zu akzeptieren sind. Die Verbraucherrechtstage 2021, veranstaltet und initiiert durch das BMJV,[6] widmen sich vor diesem Hintergrund den Herausforderungen der Künstlichen Intelligenz für das Recht in seiner ganzen Breite. Eingenommen wird dabei ein wissenschaftlich-reflektierender Blickwinkel, der zugleich Fragen der möglichen künftigen Regulierung und damit der Rechtspolitik adressiert – in dem Bestreben, auf dem Weg zu einer vertrauenswürdigen Verwendung von Künstlicher Intelligenz in Deutschland und Europa einen großen Schritt voranzukommen.

[5] Vgl. hierzu sogleich die Ausführungen unter IV.

[6] Aufgrund eines Ressortwechsels des Verbraucherschutzes in das BMUV wird dieser Tagungsband zu den Verbraucherschutztagen 2021 nunmehr vom BMUV herausgegeben.

II. Nationale KI-Regulierung im Spiegel europäischer und internationaler Regulierungsbestrebungen

Die KI-Regulierung ist keine lediglich nationale Angelegenheit. Sie ist vielmehr eingebettet in ein gesamteuropäisches Vorhaben, das zuletzt Ausdruck in einem entsprechenden Vorschlag der Europäischen Kommission gefunden hat.[7] Darin wählt die Kommission einen risikobasierten Ansatz, um anhand der Kritikalität der jeweiligen KI-Anwendung zu entscheiden, inwieweit sie rechtlich akzeptabel ist bzw. inwieweit es zu ihrer Zulässigkeit der Wahrung spezifischer rechtlicher Anforderungen bedarf.[8] Der Kommissionsvorschlag vom 21. April 2021 steht in einer Reihe von europäischen Regulierungsüberlegungen, die durch die Einrichtung einer High Level Expert Group on Artificial Intelligence[9] ihren Ausgang genommen hat und in Deutschland nicht zuletzt durch die Arbeit der Datenethikkommission[10] maßgeblich unterstützt wurde. Im Anschluss an ihre Empfehlungen wurde von der Kommission ein Whitepaper zur KI-Regulierung[11] vorgelegt, auf dem der aktuelle Vorschlag aufbaut und das er zugleich weiter konkretisiert. Der risikobasierte Ansatz unterscheidet die Stufen des unannehmbaren, des hohen, des geringen und des minimalen Risikos.[12] Während KI-Systeme mit unannehmbaren Risiken aufgrund des Verstoßes gegen die Werte der Europäischen Union und die Grundrechte verboten werden, gilt für die Systeme der übrigen Risikostufen ein gestuftes Verfahren des staatlichen Umgangs mit diesen.[13] Nicht zuletzt die Pflichten der Anbieter von KI-Systemen mit hohem Risiko werden in dem Kommissionsvorschlag klar definiert.[14] Auch Vorschläge zur nationalen Rechtsdurchsetzung werden unterbreitet.[15] Dabei betont die Kommission, dass durch

---

[7] 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 vom 21.4.2021, abrufbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (letzter Zugriff 6.11.2021); siehe dazu überblicksartig aus der deutschsprachigen Literatur z. B. *Bomhard/Merkle*, RDi 2021, 276; *Rostalski/Weiss* (Fn. 4), 329.

[8] *Valta/Vasel*, ZRP 2021, 142 (142).

[9] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (letzter Zugriff 6.11.2021).

[10] https://www.bmi.bund.de/DE/themen/it-unddigitalpolitik/datenethikkommission/datenethikkommission-node.html (letzter Zugriff 6.11.2021); diese veröffentlichte 2019 ein ausführliches Gutachten, abrufbar unter https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6 (letzter Zugriff 6.11.2021).

[11] In deutscher Sprache abrufbar unter https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf (letzter Zugriff 6.11.2021).

[12] Die Terminologie ist im deutschsprachigen Schrifttum nicht immer vollumfänglich einheitlich, siehe dazu *Hoffmann*, K&R 2021, 369 (370), *Spindler*, CR 2021, 361 (365) und *Bomhard/Merkle* (Fn. 6), 279.

[13] Dazu *Bomhard/Merkle* (Fn. 6), 279–282.

[14] Erläuternd zu diesen *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188 (1191).

[15] *Bomhard/Merkle* (Fn. 6), 282.

die nunmehr beabsichtigte verstärkte europäische Regulierung des Einsatzes von KI-Anwendungen Innovationen nicht gehemmt, sondern im Gegenteil gefördert werden: Regulierung schafft Vertrauen – und damit eine Grundvoraussetzung für die Akzeptanz von Innovationen durch die Gesellschaft, ohne die es einen entsprechenden Fortschritt nicht geben kann.[16]

Dabei finden sich Regulierungsbestrebungen im Hinblick auf Künstliche Intelligenz nicht lediglich in Europa. Auch in den USA ist das Thema bereits auf die politische Agenda gerückt. Die US-Regierung hat noch während der Amtszeit von Präsident Trump Grundsätze veröffentlicht, die Behörden bei der Ausarbeitung von Gesetzen und Regeln für den Einsatz Künstlicher Intelligenz im privaten Sektor berücksichtigen sollten.[17] Es handelt es sich um die Prinzipien des öffentlichen Vertrauens, der Transparenz, der Fairness, des Risikomanagements, des Fortschritts und der Sicherheit. Daneben hat der Kongress dem Weißen Haus die Aufgabe erteilt, eine neue Behörde zu gründen, die sämtliche dieser Schritte lenken soll (National AI Initiative Office).[18] Insgesamt erweist sich der US-amerikanische Ansatz indessen weniger regulierungsfreudig als der europäische.

III. KI-Systeme als Herausforderung für das Antidiskriminierungsrecht

Viele KI-Systeme „lernen" auf der Basis einer großen Zahl an Daten. Sowohl die Qualität der Daten – insbesondere ihre Diversität – als auch der Umgang mit ihnen durch den Menschen, an dem der Algorithmus sich orientiert, entscheiden darüber, ob die Anwendung eine diskriminierende Wirkung entfaltet,[19] die von Rechts wegen inakzeptabel ist. Nationales Antidiskriminierungsrecht findet sich derzeit im einfachen Recht vor allem im Allgemeinen Gleichbehandlungsgesetz (AGG).[20] Auch von der DS-GVO wird sich ein zumindest mittelbarer Schutz vor unzulässiger Diskriminierung versprochen.[21] Gleichwohl scheinen die Herausforderungen, die KI-Systeme unter dem Gesichtspunkt der Antidiskriminierung an das Recht stellen, derzeit nicht angemessen bewältigt. Dies gilt vor allem im Hinblick auf die Reichweite der bestehenden Vorschriften, die teilweise als zu eng

---

[16] Vgl. zur treibenden Kraft von Vertrauen die beigefügte Begründung zum KI-Verordnungsvorschlag (Fn. 6), unter 3.3.

[17] https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf (letzter Zugriff 30.11.2021).

[18] Siehe Section 102 Bill H.R. 6216, abrufbar unter https://www.congress.gov/bill/116th-congress/house-bill/6216/text (letzter Zugriff 6.11.2021).

[19] Vgl. *Wischmeyer*, AöR 143 (2018), 1 (27).

[20] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, 2019, S. 230 f.

[21] Vgl. *Hacker*, Common Market Law Review 2018, 1143, 1171 ff. sowie auf ihn Bezug nehmend *Scheer*, Algorithmen und ihr Diskriminierungsrisiko. Eine erste Bestandsaufnahme, 2019, S. 14.

eingestuft wird.[22] Daneben treten Schwierigkeiten der Rechtsdurchsetzung. Zwar spricht vieles dafür, der von einer Diskriminierung durch ein KI-System betroffenen Person die in § 22 AGG vorgesehene Darlegungs- und Beweislasterleichterung zuteilwerden zu lassen. Indessen erweist sich der Nachweis einer ungerechtfertigten Benachteiligung nach wie vor als problematisch,[23] selbst wenn der oder dem Betreffenden ein zusätzlicher Auskunftsanspruch[24] eingeräumt wird. Das Antidiskriminierungsrecht stellt vor diesem Hintergrund in Bezug auf die Künstliche Intelligenz bedeutsame Anforderungen an künftige rechtspolitische Bestrebungen.

IV. Rechtliche Anforderungen an die Transparenz und Erklärbarkeit von KI-Systemen

Im Hinblick auf KI-Systeme kann zudem zwischen den Anforderungen der Transparenz und der Erklärbarkeit unterschieden werden. Transparenz bedeutet, dass die Arbeitsprozesse der Anwendung vollständig nachvollziehbar sind bzw. gemacht werden.[25] Aus praktischer Sicht erweist sich diese Forderung in Bezug auf KI-Systeme in aller Regel als nur schwer erfüllbar. Grund hierfür ist die hohe Komplexität vieler Modelle.[26] Demgegenüber kennzeichnet der Begriff der Erklärbarkeit die Möglichkeit, für eine konkrete Einzelentscheidung der jeweiligen KI-Anwendung die wesentlichen Einflussfaktoren aufzuzeigen.[27] Aus technischer Sicht liegt die hierfür zu überwindende Hürde deutlich niedriger als diejenige im Hinblick auf die Herstellung von Transparenz.[28]

Auf ML-Techniken basierende KI-Systeme erweisen sich für ihre Anwenderin oder ihren Anwender nicht selten als „Black Box". Der Begriff umschreibt den Umstand, dass oftmals nicht nachvollziehbar ist, wie das System im Einzelfall zu seiner Entscheidung gekommen ist.[29] Ursächlich dafür ist insbesondere das „Selbstlernen" des Systems, wodurch mitunter Modifikationen und Weiterent-

---

[22] *Martini* (Fn. 19), S. 235.

[23] Im Kontext mit § 22 AGG sieht *Martini* (Fn. 19), S. 247 in dem Umstand, dass nach dieser Vorschrift immer noch „Indizien" für eine Diskriminierung vorgelegt werden müssen, eine Hürde.

[24] Ein (begrenzter) Auskunftsanspruch hinsichtlich der Nachvollziehbarkeit einer KI-Anwendung kann sich aus der DS-GVO ergeben, siehe dazu unten unter V.; ob ein etwaiger Auskunftsanspruch gegen den Anspruchsgegner auf Basis des AGG besteht, ist umstritten, vgl. *Thüsing*, in: MüKo-BGB, Bd. 1, 9. Aufl. 2021, § 22 AGG Rn. 8.

[25] *Huber/Giesecke*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik. Rechtshandbuch, 2020, § 19 Rn. 36.

[26] Zu diesem Problem siehe *Käde/von Maltzan*, CR 2020, 66 (67).

[27] *Huber/Giesecke* (Fn. 24), § 19 Rn. 36.

[28] *Döbel* et al., Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung, S. 30, abrufbar unter https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf (letzter Zugriff 6.11.2021).

[29] *Ebers*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik. Rechtshandbuch, 2020, § 3 Rn. 25.

wicklungen des erlernten Entscheidungsmodells erfolgen, die selbst für die Entwicklerin oder den Entwickler nicht immer vorhersehbar sind.[30] Zudem sind die kausalen Ursachen für die eine oder andere Entscheidung, die auf der Basis des erlernten Modells getroffen wird, auch im Nachhinein oftmals nicht restlos aufzuklären.[31] Indessen spielen Gründe für die normative Bewertung, die das Recht vornimmt, eine entscheidende Rolle. Sofern beispielsweise eine Kreditvergabe auf Empfehlung eines KI-Systems hin abgelehnt wird, stellt sich die Frage, ob dies zu Recht erfolgt ist. Mangelnde Transparenz bzw. Erklärbarkeit kann allerdings dazu führen, dass hierauf letztlich keine Antwort gegeben werden kann. Aufgrund der teilweise hohen Eingriffsintensität von KI-Systemen in Bezug auf Rechte von Verbraucherinnen und Verbraucher erweist sich ein dementsprechendes Maß an Erklärbarkeit als Desiderat, das auch rechtlich zu verlangen ist. Dies zeigt sich nicht zuletzt sowohl im zivilrechtlichen als auch im strafrechtlichen Kontext im Hinblick auf den darin oftmals zu führenden Nachweis der Kausalität, der durch die Intransparenz von KI-Systemen erheblich erschwert bzw. in Gänze unmöglich gemacht wird.[32] Reformbestrebungen können sich insoweit auf die Sicherstellung eines sachgerechten Umfangs an Erklärbarkeit richten, der sich an der Kritikalität der jeweiligen KI-Anwendung orientiert.[33] Dabei können Anforderungen an die Transparenz und Erklärbarkeit an unterschiedlichen Stellen ansetzen – etwa am Algorithmus selbst oder aber am Design der KI-Anwendung. Insoweit beschreibt „Transparency by design" die Forderung nach einer technischen Entwicklung von KI-Anwendungen, die intrinsisch erklärend sind.[34]

V. Rechtliche Anforderungen an die Nachvollziehbarkeit und Überprüfbarkeit von KI-Systemen

Transparenz und Erklärbarkeit stehen in unmittelbarem Zusammenhang mit dem Interesse an Nachvollziehbarkeit und Überprüfbarkeit von KI-Systemen. Vollständige Transparenz stellt in jedem Fall Nachvollziehbarkeit her. Indessen ist Trans-

---

[30] Vgl. zum „Weiterlernen" *Konertz/Schönhof*, Das technische Phänomen „Künstliche Intelligenz" im allgemeinen Zivilrecht. Eine kritische Betrachtung im Lichte von Autonomie, Determinismus und Vorhersehbarkeit, 2020, S. 49 ff.

[31] *Martini* (Fn. 19), S. 43.

[32] Zur Relevanz der Intransparenz von KI bei der Kausalitätsbewertung im Strafrecht siehe *Lohmann*, Strafrecht im Zeitalter von Künstlicher Intelligenz. Der Einfluss von autonomen Systemen und KI auf die tradierten strafrechtlichen Verantwortungsstrukturen, 2021, S. 156 f.; zur Kausalitätsproblematik im Zivilrecht *Haagen*, Verantwortung für Künstliche Intelligenz. Ethische Aspekte und zivilrechtliche Anforderungen bei der Herstellung von KI-Systemen, 2021, S. 327.

[33] So in gewisser Weise im KI-VO-E der EU-Kommission vom April 2021 angelegt: Siehe Art. 13 Abs. 1 S. 1 KI-VO-E und Art. 14 KI-VO-E, die Anforderungen (nur) für „Hochrisiko-KI-Systeme" statuieren, dazu *Ebert/Spiecker gen. Döhmann* (Fn. 13), 1190.

[34] Als Überblick: *Felzmann/Fosch-Villaronga/Lutz/Tamò-Larrieux*, Science and Engineering Ethics 2020, 3333.

parenz im Einzelfall nicht erforderlich, um die Entscheidung eines KI-Systems nachvollziehbar zu machen. Dabei stellt die Nachvollziehbarkeit einen relevanten Faktor im Kontext des Schutzes von Verbraucherrechten dar. Sie ist Grundvoraussetzung für die Überprüfbarkeit von Entscheidungen, die unter Verwendung eines KI-Systems getroffen worden sind. Zur Erlangung von Kenntnissen, die auf die Nachvollziehbarkeit der jeweiligen Entscheidung gerichtet sind, kann der datenschutzrechtliche Auskunftsanspruch gemäß Art. 13 Abs. 2 lit. f i. V. m. Art. 15 Abs. 1 lit. h DS-GVO herangezogen werden. Allerdings kann dieser mit dem rechtlichen Schutz von Geschäftsgeheimnissen i. S. v. § 2 Nr. 1 GeschGehG in Konflikt geraten.[35] Grundsätzlich kann das Auskunftsverlangen nur für solche Daten verweigert werden, bei denen das Geheimhaltungsinteresse als Recht eines Dritten nach Art. 15 Abs. 4 DS-GVO das Auskunftsrecht überwiegt.[36] Auf einer Linie damit liegt Erwägungsgrund 63 S. 5 zur DS-GVO, wonach das Auskunftsrecht keine Geschäftsgeheimnisse beeinträchtigen soll.[37] Für KI-Anwendungen lässt sich hieraus ableiten, dass gegenüber der oder dem Auskunftsberechtigten lediglich die Entscheidungsmechanismen dargelegt werden müssen, nicht hingegen die exakten Berechnungsformeln.[38] Diese Annahme stützt auch eine Entscheidung des BGH aus dem Jahr 2014,[39] wonach die Schufa nach § 34 BDSG a. F. nicht verpflichtet war, über die genaue Scorewertberechnung mit den eingeflossenen Rechengrößen Auskunft zu erteilen, da auch Rechengrößen als Geschäftsgeheimnis einzustufen seien. Indessen kann sich dies als erhebliches Hindernis im Hinblick auf die Rechtsdurchsetzung von Verbraucherinnen und Verbrauchern erweisen, weshalb fraglich ist, ob künftige KI-Regulierung insoweit auf eine Stärkung von Auskunftsverlangen – notfalls durch prozessuale Instrumente – gerichtet sein sollte.

VI. Verbraucherschutz durch „gute" Trainingsdaten

Die Qualität der Trainingsdaten entscheidet über die tatsächliche und rechtliche Qualität des gesamten KI-Systems. Werden die Trainingsdaten beispielsweise lediglich einseitig ausgewählt, kann dies zu Verzerrungen führen, die im Zweifel sogar diskriminierende Folgen nach sich ziehen.[40] Ein Beispiel liefern KI-Anwendungen, die im Kontext von Bewerbungsverfahren zum Einsatz kommen. Sofern

---

[35] *Thüsing/Rombey*, ZD 2020, 221 (222).

[36] Vgl. *Thüsing/Rombey* (Fn. 34), 222.

[37] Siehe zu diesem auch *Thüsing/Rombey* (Fn. 34), 222.

[38] *Martini* (Fn. 19), S. 181, 342; vgl. zu der komplexen Frage, welche Informationen bei ML-Systemen anzugeben sind, *Sesing*, MMR 2021, 288 (291 m. w. N.).

[39] BGH, Urt. v. 28.1.2014 – VI ZR 156/13.

[40] *Langer/Weyerer*, in: Oswald/Borucki (Hrsg.), Demokratietheorie im Zeitalter der Frühdigitalisierung, 2020, 219 (224 f.).

die Trainingsdaten ein Bias zulasten von Frauen und nichtbinären Personen aufweisen, wird sich dies unweigerlich diskriminierend auf die durch das System empfohlene Auswahlentscheidung auswirken.[41] Hiermit gehen hohe Gefahren einher – insbesondere dann, wenn die für die Empfehlung ursächlichen Faktoren für die Systemanwenderin oder den Systemanwender intransparent sind. Die Qualität von Trainingsdaten hängt außerdem davon ab, inwieweit sie rechtliche Vorgaben des Datenschutzes wahren. Sensible, personenbezogene Daten dürfen lediglich unter Wahrung der gesetzlichen Bestimmungen zum Einsatz kommen. Zu berücksichtigen ist dabei insbesondere Art. 9 DS-GVO, wonach die Verarbeitung besonderer Kategorien personenbezogener Daten, etwa die ethnische Herkunft oder die politische Meinung, grundsätzlich untersagt ist.[42]

Datenqualität dient damit zugleich dem Schutz der Rechtsposition des Einzelnen. Aus rechtlicher Perspektive ist daher zunächst zu definieren, was Daten zu „guten" macht. Der normative Blickwinkel stößt allerdings unter Umständen in bestimmten Bereichen an die Grenzen des real Umsetzbaren. Nicht immer sind Daten in der rechtlich wünschenswerten Qualität in hinreichender Zahl vorhanden. Datenanonymisierung führt nicht selten dazu, dass viele für den Lernprozess relevante Daten verlorengehen.[43] Selbst synthetische Daten[44] vermögen diese Lücke regelmäßig nicht zu schließen, zumal ein sinnvolles Maschinelles Lernen voraussetzt, dass die synthetischen Daten mit den Ausgangsdaten eine hinreichende Übereinstimmung aufweisen.[45] Damit ist das identische Problem des Datenschutzes lediglich in einem „anderen Gewande" erneut gegeben.[46] Auch diese Schwierigkeit ist normativ zu beantworten: Soll bei minderer Datenqualität unter Umständen in Gänze auf eine KI-Anwendung verzichtet werden oder nehmen wir dies zugunsten der Anwendbarkeit eines bestimmten KI-Systems hin? Eine mögliche technische Antwort kann im Verfahren des Föderalen Lernens liegen.[47] Darin findet das Training der KI dezentral statt, etwa auf Smartphones

---

[41] Für ein Beispiel der Diskriminierung von Frauen siehe *Verhoeven*, in: Verhoeven (Hrsg.), Digitalisierung im Recruiting. Wie sich Recruiting durch künstliche Intelligenz, Algorithmen und Bots verändert, 2020, 225 (237).

[42] *Hacker* (Fn. 20), 1182; *Scheer* (Fn. 20), S. 14. Zu Art. 9 DS-GVO im Machine-Learning-Kontext siehe auch *Niemann/Kevekordes*, CR 2020, 179 (179 ff.).

[43] *Kaulartz*, in: Kaulartz/Braegelmann (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 8.9 Rn. 20.

[44] *Hillmer*, Daten als Rohstoffe und Entwicklungstreiber für selbstlernende Systeme. Zum Regulierungsbedürfnis von Innovationshemmnissen durch Datennetzwerkeffekte, 2021, 242 definiert diese als „computergenerierte Daten, die empirische Daten nachahmen."

[45] *Kaulartz* (Fn. 42), Kap. 8.9 Rn. 22.

[46] *Kaulartz*, Datenschutz-Compliance bei KI am Beispiel Federated Learning, 18.11.2019, abrufbar unter https://www.cmshs-bloggt.de/tmc/machine-learning-datenschutz-compliance-bei-ki-am-beispiel-federated- learning/ (letzter Zugriff 6.11.2021).

[47] *Kaulartz* (Fn. 45); *Huth*, in: Kaulartz/Braegelmann (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 2.3 Rn. 12 ff.

verschiedener Personen.[48] Die lokalen Modelle werden sodann in einem übergeordneten Modell zusammengeführt, das die dezentral gesammelten Informationen infolge einer Verschlüsselung nicht erfährt.[49] Der Vorteil des Verfahrens liegt darin, dass die personenbezogenen Daten zu keinem Zeitpunkt vom dezentralen Sammelort getrennt und stets nur die infolge der Anonymisierung datenschutzrechtlich unbedenklichen Modelle übermittelt werden.[50] Es bleibt allerdings offen, ob sich damit sämtliche datenschutzrechtlichen Probleme abschließend lösen lassen.

VII. Verbraucherschutz durch Normen, Standards und Zertifizierung von KI-Systemen

Für den Alltag von Verbraucherinnen und Verbrauchern halten KI-Systeme in großem Umfang Erleichterungen und Verbesserungen bereit. Diese Chancen werden allerdings durch Risiken flankiert, die in der Wahrnehmung vieler nach wie vor dominieren und auf diese Weise ein Misstrauen gegenüber der Technologie begründen. Ein solches Misstrauen kann allerdings dazu führen, dass selbst für Verbraucherinnen und Verbraucher besonders vorteilhafte KI-Anwendungen keinen Einzug in ihre Lebenswirklichkeit halten können – es erweist sich damit als Hemmnis für sowohl den Fortschritt als auch die Verbesserung von Lebensumständen der Verbraucherinnen und Verbraucher.

Dieser Entwicklung können Normen, Standards[51] sowie Verfahren der Zertifizierung entgegenwirken.[52] Normen und Standards schaffen klare Vorgaben, die bei der Entwicklung von KI-Systemen und ihrem Einsatz Berücksichtigung finden müssen. Ihre Einhaltung schafft Vertrauen, indem die Verbraucherin oder der Verbraucher eine klare Vorstellung von dem jeweiligen KI-System erlangt, mit dem sie oder er sich umgibt. Zertifikate ermöglichen dabei eine Selbstermächtigung von Verbraucherinnen und Verbrauchern im Hinblick auf das zertifizierte Produkt, indem sie für einen bestimmten Qualitätsstandard in Bezug auf Krite-

---

[48] *Kaulartz* (Fn. 45).

[49] *Kaulartz* (Fn. 45).

[50] *Kaulartz* (Fn. 45).

[51] Die Begriffe Normen und Standards werden insbesondere im Kontext der Arbeit des Deutschen Instituts für Normung (DIN) verwendet. Dieses beschreibt im Rahmen seiner Arbeit eine Norm als „Dokument, das Anforderungen an Produkte, Dienstleistungen oder Verfahren festlegt", wobei DIN-Normen in einer speziellen Vorgehensweise entwickelt werden. Als (bloße) Standards werden beim DIN inhaltlich entsprechende Festlegungen bezeichnet, die nicht das umfangreiche Prozedere für die Erzeugung von DIN-Normen durchlaufen haben. Siehe für die gesamte Fußnote https://www.din.de/de/ueber-normen-und-standards/basiswissen (letzter Zugriff 21.1.2022).

[52] Zur Relevanz von „Zertifizierungen und Standards" beim Umgang mit KI siehe auch die *Datenethikkommission*, Empfehlungen der Datenethikkommission für die Strategie Künstliche Intelligenz der Bundesregierung, S. 3, abrufbar unter https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/themen/it-digitalpolitik/datenethikkommission/empfehlungen-datenethikkommission.pdf?__blob=publicationFile&v=2 (letzter Zugriff 6.11.2021).

rien wie den Datenschutz, die Transparenz und Erklärbarkeit bzw. Sicherheit des KI-Systems bürgen.

Erste Schritte in Richtung einer Regulierung im Bereich von Normen, Standards und Zertifizierung sind national bereits gegangen worden. In der Deutschen Normungsroadmap Künstliche Intelligenz[53] erfolgt beispielsweise eine umfangreiche Analyse des Bestandes und Bedarfs an internationalen Normen und Standards für die KI-Technologie. Umfasst sind davon neben technischen auch ethische und gesellschaftliche Anforderungen an Normen für KI. Zudem bestehen mehrere nationale Zertifizierungsinitiativen,[54] die in den nächsten Jahren konkrete Ergebnisse erwarten lassen. Die Entwicklung steht gleichwohl noch am Anfang. Es lohnt daher, insbesondere aus Sicht des Verbraucherschutzes einen (nicht zuletzt rechtspolitischen) Blick auf die notwendigen Anforderungen an Normen, Standards und Verfahren der Zertifizierung zu werfen. Dabei geht es gerade auch darum, zu fragen, wie die Interessen von Verbraucherinnen und Verbrauchern in Verfahren der Standardisierung angemessene Berücksichtigung finden können.

VIII. Exemplarische Verdeutlichung der Herausforderungen von KI-Systemen an das Recht

Die bislang abstrakt beschriebenen Herausforderungen, die KI-Systeme an die Rechtsordnung stellen, zeigen sich in einer Vielzahl unterschiedlicher Lebensbereiche. Die Bewältigung der damit einhergehenden regulatorischen Aufgaben erweist sich als besonders bedeutsam im Hinblick auf diejenigen KI-Anwendungen, die mit erheblicher Intensität in die Rechte von Verbraucherinnen und Verbrauchern eingreifen. Hierzu gehört zum Beispiel der Einsatz von KI innerhalb der Finanzdienstleistung, des Versicherungswesens, des Arbeitsmarkts, der Mobilität, der Medizin und des Zugangs zum Recht.

1. KI-Systeme in der Finanzdienstleistung

KI-Systeme prägen zunehmend die Abläufe der Finanzdienstleistung. Zu denken ist etwa an Verfahren des Kreditscorings und des Robo-Advisings. Kreditscoring umschreibt eine algorithmenbasierte Kreditwürdigkeitsprüfung.[55] Hiermit

---

[53] *DIN/DKE*, Deutsche Normungsroadmap Künstliche Intelligenz, abrufbar unter https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf (letzter Zugriff 6.11.2021).

[54] Zu nennen sind hier das Projekt „Zertifizierte KI", https://www.zertifizierte-ki.de (letzter Zugriff 6.11.2021) und das Projekt „ExamAI – KI-Testing & Auditing", https://www.rechtsinformatik.saarland/de/forschung/projekte/examai (letzter Zugriff 6.11.2021).

[55] Vgl. *Helfrich*, Kreditscoring und Scorewertbildung der SCHUFA. Datenschutzrechtliche Zulässigkeit im Rahmen der praktischen Anwendung, 2010, S. 22.

gehen prinzipiell Risiken für Rechte von Verbraucherinnen und Verbrauchern einher. Fehlentscheidungen können erhebliche Folgen für die Einzelne oder den Einzelnen haben. Dabei können Fehler insbesondere daraus erwachsen, dass die Trainingsdaten der KI-Anwendung mit einem Bias behaftet sind, was zu ungerechtfertigten Diskriminierungen Einzelner führen kann. Rechtliche Rahmenbedingungen für das Kreditscoring ergeben sich insbesondere aus § 31 Abs. 1 Nr. 2 BDSG, wonach Scoring nur dann zulässig ist, wenn „die zur Berechnung des Wahrscheinlichkeitswerts genutzten Daten unter Zugrundelegung eines wissenschaftlich anerkannten mathematisch-statistischen Verfahrens nachweisbar für die Berechnung des bestimmten Verhaltens erheblich sind". Problematisch hieran erscheint insbesondere, dass das Wissenschaftlichkeitserfordernis gerade keinen Qualitätsstandard etabliert.[56] Im Kontext des Kreditscorings werden Forderungen nach strengeren Regularien und Maßstäben für das Prognosesystem immer lauter.[57] Um sachgerecht vor Diskriminierung zu schützen, sind außerdem Auskunfts- und Verständlichkeitsansprüche von Verbraucherinnen und Verbrauchern, etwa nach der DS-GVO, ein weiteres Desiderat.[58]

Robo-Advising umschreibt die algorithmengesteuerte Erstellung eines Portfolios entsprechend der gewünschten Anlagestrategie.[59] Darüber hinaus können davon der Erwerb und die Strukturierung eines Portfolios und seine fortlaufende Umschichtung erfasst sein.[60] Derzeit ist Robo-Advising allein durch Finanzdienstleister, nicht unmittelbar durch Anlegerinnen und Anleger nutzbar.[61] Risiken für den Verbraucherschutz liegen dabei insbesondere in dem häufig bestehenden Informationsdefizit im Hinblick auf die Funktionsweise des Robo-Advisors.[62] Der Wunsch nach Transparenz und damit der Offenlegung des Algorithmus kann allerdings mit dem darauf gerichteten Geheimhaltungsinteresse der Betreiberin oder des Betreibers in Konflikt geraten. Nicht zuletzt ist sicherzustellen, dass Robo-Advisors den Verhaltensregeln der §§ 63 ff. WpHG entsprechen.[63] Die Überprüfung dieser Anforderung wirft wiederum praktische Schwierigkeiten auf, denen ggf. durch rechtliche Regulierung begegnet werden muss.

---

[56] *Gerberding/Wagner*, ZRP 2019, 116 (118).

[57] *Gerberding/Wagner* (Fn. 55), 119.

[58] *SVRV*, Verbrauchergerechtes Scoring, Gutachten des Sachverständigenrats für Verbraucherfragen, S. 4 f., 143, abrufbar unter https://www.svr-verbraucherfragen.de/wp-content/uploads/SVRV_Verbrauchergerechtes_Scoring.pdf (letzter Zugriff 6.11.2021).

[59] *Denga*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik. Rechtshandbuch, 2020, § 15 Rn. 3, 8 ff.

[60] *Denga* (Fn. 58), § 15 Rn. 13; für weitere potenzielle Aufgaben von KI-Technik im „Assetmanagement" siehe *Voß*, in: Kaulartz/Braegelmann (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 13.2 Rn. 4.

[61] *Denga* (Fn. 58), § 15 Rn. 14.

[62] *Denga* (Fn. 58), § 15 Rn. 7.

[63] *Denga* (Fn. 58), § 15 Rn. 33 ff.; siehe auch *Voß* (Fn. 59), Kap. 13.2 Rn. 13.

2. KI-Systeme im Versicherungswesen

Im Versicherungswesen finden KI-Systeme bereits in vielfacher Weise Verwendung. In den Vordergrund rückt dabei zunächst solche algorithmengesteuerte Software, die zur Einschätzung von Risiken herangezogen wird, auf deren Basis eine automatische Prämienkalkulation erfolgt.[64] Daneben treten Verfahren des „Smart Underwriting". Diese umfassen zum Beispiel die Auswahl von Fragen für einen Vertragsabschluss – etwa basierend auf einem Foto des zu versichernden Objekts – sowie die sofortige Angebotserstellung unter Rückgriff auf bereits von der Kundin oder dem Kunden gespeicherte Daten („Smart Questionnaire").[65] Nicht zuletzt wird KI in diesem Zusammenhang zur Minimierung von Betrugsgefahren durch den Abgleich von Daten und durch Plausibilitätskontrollen eingesetzt.[66] Zu denken wäre dabei etwa an die Heranziehung von Schadensbildern, der Schilderung des Schadenshergangs sowie von Auffälligkeiten der Kundin oder des Kunden in der Vergangenheit. Teilweise kommt es dabei zur Einbeziehung externer Datenbanken, etwa des Wetterdienstes etc.[67] Nicht zuletzt wird KI im Versicherungswesen zur Kommunikationsverwaltung verwendet. Hier dient die algorithmenbasierte Software zum Beispiel der Kategorisierung von Kundenanfragen, der Beratung durch Chatbots bei einfach gelagerten Fragestellungen oder der Schadensmeldung.[68]

Aus rechtlicher Perspektive wirft der beschriebene Einsatz von KI im Versicherungswesen eine Vielzahl an Fragen auf. Diese betreffen zum einen den Datenschutz. Kommt es zum Austausch hochsensibler Daten, ist eine Anonymisierung erforderlich. Damit keine Fehler zulasten der Kundin oder des Kunden auftreten, muss außerdem eine durchgehend hohe Datenqualität sichergestellt werden. Nur so können die Daten umfassend ausgewertet werden. Zum anderen ist unter dem Blickwinkel der Diskriminierungsverbote des AGG (z. B. § 20 Abs. 2 AGG) darauf zu achten, dass die KI darauf trainiert wird, bestimmte Daten unberücksichtigt zu lassen, auch wenn diese tatsächlich risikoerhöhend sind.[69] Weil die konkrete Höhe der errechneten Prämie für die Kundin oder den Kunden von erheblicher

---

[64] Vgl. *Mühlenbruch*, Datenbasierte Risikoberechnung: Wie Künstliche Intelligenz Versicherern einen kräftigen Schub verleiht, 19.7.2021, https://versicherungswirtschaft-heute.de/maerkte-und-vertrieb/2021-07-19/datenbasierte-risikoberechnung-wie-kuenstliche-intelligenz-versicherern-einen-kraeftigen-schub-verleiht/ (letzter Zugriff 24.1.2022).

[65] *Ebert*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik. Rechtshandbuch, 2020, § 16 Rn. 12.

[66] *Ebert* (Fn. 64), § 16 Rn. 21; vgl. auch Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Dr. Florian Toncar, Christian Dürr, Frank Schäffler, weiterer Abgeordneter und der Fraktion der FDP – BT-Drucksache 19/5943, S. 6: „Im Versicherungssektor erscheinen die Möglichkeiten im Bereich Betrugserkennung […] besonders aussichtsreich."

[67] *BaFin*, Big Data trifft auf künstliche Intelligenz, S. 112, abrufbar unter https://www.bafin.de/SharedDocs/Downloads/DE/dl_bdai_studie.html (letzter Zugriff 6.11.2021).

[68] *Ebert* (Fn. 64), § 16 Rn. 20.

[69] *Ebert* (Fn. 64), § 16 Rn. 10.

Relevanz ist, stellen sich auch insoweit Fragen nach Verbraucherrechten im Hinblick auf die Nachvollziehbarkeit und Überprüfbarkeit des zugrunde liegenden Algorithmus. In Bezug auf „Smart Underwriting"-Verfahren müssen außerdem die Vorgaben von Art. 22 DS-GVO und § 37 BDSG hinsichtlich automatisierter Entscheidungen gewahrt werden.[70] Hier können insbesondere Fragen danach auftreten, ab welchem Punkt der Vorgang an eine menschliche Sachbearbeiterin oder einen menschlichen Sachbearbeiter abgegeben werden muss bzw. welche Verträge ausschließlich elektronisch geschlossen werden dürfen.[71] Nicht zuletzt stellen algorithmenbasierte Systeme im Versicherungswesen die Versicherungsaufsicht (§§ 43, 294 VAG i. V. m. Art. 22 DS-GVO) vor Herausforderungen: Zu klären ist etwa, ob sich diese künftig auch auf die Berechnungsformel des verwendeten Algorithmus bzw. den Lernprozess erstreckt.[72]

3. KI-Systeme und Arbeit

KI-Systeme spielen immer häufiger eine Rolle beim Zugang von Verbraucherinnen und Verbrauchern zum Arbeitsmarkt. Eingesetzt werden sie sowohl durch Akteurinnen und Akteure in der Wirtschaft, die beispielsweise Auswahlentscheidungen in Bewerbungsverfahren auf der Basis von Empfehlungen treffen, die ihnen ein KI-System liefert.[73] Hinzu treten KI-basierte Mitarbeiteranalysen,[74] die für innerbetriebliche Entscheidungsprozesse herangezogen werden. Doch auch im öffentlichen Sektor ist die Verwendung von Entscheidungsunterstützungssoftware keine Seltenheit mehr. Ein Beispiel liefert eine KI-Software, die der österreichische Arbeitsmarktservice einsetzt, um Arbeitsmarktchancen arbeitsloser Personen zu bewerten.[75] Die Empfehlungen des KI-Systems fließen zum Beispiel

---

[70] Nach Art. 22 Abs. 1 DS-GVO verfügt jede „betroffene Person" über „das Recht, nicht einer ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt." Dies gilt allerdings nicht unbeschränkt (Art. 22 Abs. 2 DS-GVO, bei dem seinerseits die Abs. 3 und 4 zu beachten sind), dazu *Martini*, in: Paal/Pauly (Hrsg.), Datenschutz-Grundverordnung, Bundesdatenschutzgesetz, 3. Auflage 2021, Art. 22 DSGVO Rn. 30 u. 30a. § 37 BDSG enthält – entsprechend der Regelungsoption für die Mitgliedstaaten in Art. 22 Abs. 2 lit. b DS-GVO – zusätzliche Konstellationen, in denen eine automatisierte Entscheidung entgegen der Grundaussage des Art. 22 Abs. 1 DS-GVO zulässig sind, *Helfrich*, in: Sydow (Hrsg.), Bundesdatenschutzgesetz, 2020, § 37 BDSG Rn. 1 u. 2.

[71] Vgl. *Ebert* (Fn. 64), § 16 Rn. 23.

[72] *Armbrüster/Prill*, ZfV 2020, 110 (112 f.).

[73] *Verhoeven*, in: Verhoeven (Hrsg.), Digitalisierung im Recruiting. Wie sich Recruiting durch künstliche Intelligenz, Algorithmen und Bots verändert, 2020, 113 (120). Vgl. für KI-Nutzungen bei der Einstellung auch *Hinz*, in: Kaulartz/Braegelmann (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 11 Rn. 1 ff.

[74] Ein Beispiel für eine Mitarbeiteranalyse findet sich in *Hinz* (Fn. 72), Kap. 11 Rn. 2.

[75] Siehe *Schäfer*, in: Kipker/Voskamp (Hrsg.), Sozialdatenschutz in der Praxis, 2021, Kap. 7 Rn. 20.

bei der Frage ein, ob eine arbeitslose Person seitens des Arbeitsmarktservice eine Finanzierung für eine Weiterbildung erhält.[76] KI-Systeme spielen im Bereich der Arbeit aus rechtlicher Perspektive insbesondere hinsichtlich der Frage der Vermeidung ungerechtfertigter Diskriminierung und der Haftung für Fehler des Systems eine Rolle.[77] Das in Österreich durch den Arbeitsmarktservice zum Einsatz kommende algorithmische Entscheidungssystem bezieht diverse Daten ein, wie etwa berufliche und solche, die die Ausbildung und bisherige Anstellung betreffen. Auch wird berücksichtigt, wie alt die Person ist, welchem Geschlecht und welcher Nationalität sie angehört.[78] Die Verwendung des Algorithmus ist geeignet, erheblichen Einfluss auf die Entscheidungen von Mitarbeiterinnen und Mitarbeitern des Arbeitsmarktservice zu nehmen.[79] Insoweit besteht nicht zuletzt das Risiko, dass durch das System getroffene Vorschläge unhinterfragt übernommen werden – selbst wenn diese zum Beispiel diskriminierendes Potenzial aufweisen.[80]

4. KI-Systeme und Mobilität

Der breiten Öffentlichkeit sind KI-Systeme oftmals in erster Linie aus dem Kontext des autonomen Fahrens bekannt. Darüber hinaus kann sich KI allerdings als Schlüsseltechnologie in der Mobilität der Zukunft erweisen. Neben zunehmend autonom gesteuerte Fahrzeuge tritt dabei eine Infrastruktur mit erhöhtem Vernetzungsgrad und selbstlernenden Systemen, die Verkehrsflüsse optimieren und zur Verkehrssicherheit beitragen.[81] Eine Rolle spielen dabei zum Beispiel „intelligente" Sensoriksysteme, die Fahrzeuge miteinander und mit ihrer übrigen Umwelt vernetzen.[82] Kommt es aber im Zusammenhang mit dem Einsatz von KI-Systemen im Bereich der Mobilität zu Verletzungen von Personen, wirft dies

---

[76] *Allhutter/Fischer/Mager*, AMS-Algorithmus am Prüfstand, ITA-Dossier Nr. 43, Institut für Technikfolgen-Abschätzung, Österreichische Akademie der Wissenschaften, S. 1, abrufbar unter http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier043.pdf (letzter Zugriff 6.11.2021).

[77] Bezüglich der Themen Arbeits- und Gesundheitsschutz sowie der Haftungsfrage siehe *Günther/Böglmüller*, BB 2017, 53 (53 ff.). Zum Thema der Diskriminierung siehe z. B. *Hinz* (Fn. 72), Kap. 11 Rn. 33 ff.

[78] *Fanta*, Österreichs Jobcenter richten künftig mit Hilfe von Software über Arbeitslose, 13.10.2018, https://netzpolitik.org/2018/oesterreichs-jobcenter-richten-kuenftig-mit-hilfe-von-software-ueber-arbeitslose/ (letzter Zugriff 6.11.2021); siehe auch *Büchner/Dosdall*, Kölner Zeitschrift für Soziologie und Sozialpsychologie 2021, 333 (338).

[79] Vgl. *Mager/Allhutter*, Wie fair ist der AMS-Algorithmus?, ITA-Dossier Nr. 52, Institut für Technikfolgen-Abschätzung, Österreichische Akademie der Wissenschaften, S. 2, abrufbar unter http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier052.pdf (letzter Zugriff 6.11.2021) und *Fanta* (Fn. 77).

[80] Vgl. *Mager/Allhutter* (Fn. 78), 2 und *Fanta* (Fn. 77).

[81] *Guggenberger*, NVwZ 2019, 844 (846).

[82] Ein Beispiel ist in diesem Kontext das „OTS 1.0"-Projekt, das durch das Bundesministerium für Umwelt, Naturschutz und nukleare Sicherheit unterstützt wird. Siehe dazu https://www.iav.com/news/siemens-campus-muenchen-perlach-wird-testfeld-fuer-autonomes-fahren/ (letzter Zugriff 6.11.2021).

aus rechtlicher Sicht nicht zuletzt[83] Fragen der zivilrechtlichen Haftung auf. Wer ist haftbar – Mensch oder Maschine? Während die einen Konzepte der e-Person diskutieren,[84] denken andere über eine Modifikation des Verantwortungsbegriffs nach, die es auch zuließe, den Menschen für Ereignisse haftbar zu machen, die er – beispielsweise aufgrund des Black Box-Charakters eines KI-Systems – gerade nicht vorhersehen konnte.[85] In Abhängigkeit davon, wie diese Fragen beantwortet werden, ist auch darüber zu entscheiden, welche Autonomiegrade von Fahrzeugen als zulässig beurteilt werden.[86]

5. KI-Systeme in der Medizin

Ein großes Anwendungsfeld für KI-Systeme liefert die Medizin. KI-Systeme dienen hierbei als Entscheidungsunterstützungssysteme, zum Beispiel bei der Früherkennung von Krankheiten.[87] Neben der Diagnose kommen sie auch in der Behandlung von Patientinnen und Patienten zum Einsatz.[88] KI-Systeme ermöglichen eine höhere Personalisierung von Behandlungsplänen und sogar spezifischen Behandlungsmethoden, etwa Verfahren der Genomeditierung[89].[90] Neben Fragen der Verantwortung und des Datenschutzes stellt sich aus rechtlicher Perspektive zudem das Problem, wie menschliche Fähigkeiten aufrechterhalten und verbessert werden können, sofern die Praxis zunehmend durch technische Systeme geprägt ist. Wenn auf KI basierende Diagnosewerkzeuge teilweise mit einer höheren Wahrscheinlichkeit ein richtiges Ergebnis erzielen als der Mensch, besteht das Risiko, dass ihr Einsatz über die Dauer der Zeit dazu führt, dass der Mensch seine eigenen Fähigkeiten immer mehr verliert – etwa weil er sich zunehmend „blind" auf die Technik verlässt und damit praktische Anwendungsmög-

---

[83] Für einen Überblick zu Fragen und Problemen der strafrechtlichen Verantwortlichkeit vgl. *Beck*, in: Ebers/Heinze/Krügel/Steinrötter (Hrsg.), Künstliche Intelligenz und Robotik. Rechtshandbuch, 2020, § 7 Rn. 11 ff.

[84] Ein Überblick hierzu findet sich bei *Ringlage*, Haftungskonzepte für autonomes Fahren – „ePerson" und „RmbH"?, 2021, S. 162 ff.

[85] Zum Thema der Gefährdungshaftung für KI *Haagen* (Fn. 31), S. 358 ff.

[86] Hinsichtlich aktueller Änderungen im StVG durch das „Gesetz zum autonomen Fahren" *Ternig*, ZfS 2021, 604 sowie *Hilgendorf*, JZ 2021, 444 (zum Gesetzesentwurf).

[87] Für Depressionen siehe *Topol*, Deep Medicine. Künstliche Intelligenz in der Medizin, 2020, S. 159 f.

[88] *Wischmann/Rohde*, in: Wittpahl (Hrsg.), Künstliche Intelligenz. Technologie | Anwendung | Gesellschaft, 2019, 99 (111 ff.).

[89] Darunter versteht man Prozeduren, die es erlauben, präselektierte Stellen eines Genoms planmäßig zu adaptieren, siehe *Deutscher Ethikrat*, Eingriffe in die menschliche Keimbahn. Stellungnahme, 2019, S. 59.

[90] *König* et al., Künstliche Intelligenz in der genomischen Medizin – Potentiale und Handlungsbedarf, abrufbar unter https://www.isi.fraunhofer.de/content/dam/isi/dokumente/cct/2021/Policy%20Brief%202021_KI%20in%20der%20genomischen%20Medizin.pdf (letzter Zugriff 6.11.2021).

lichkeiten verstreichen lässt.[91] Gesellschaftlich erweist sich dies nicht zuletzt aus dem Grund als problematisch, als Fortschritt darauf fußt, dass Menschen ihre eigenen Fähigkeiten immer weiter verbessern. Darüber hinaus stellen neue, durch KI-Systeme erlangte Möglichkeiten medizinischer Verfahren, etwa die Genomeditierung, bestehende Gesetze infrage. Diese weisen mitunter Lücken auf, da entsprechende Technologien bislang nicht bedacht wurden. Sofern sie bereits erfasst und zum Beispiel verboten sind, stellt sich angesichts des Fortschritts die Frage nach der verbleibenden Berechtigung entsprechender Normen.

6. Einsatz von KI-Systemen in der Justiz

Der Zugang zum Recht ist nicht selten durch Hürden verstellt, die Verbraucherinnen und Verbraucher von einer effektiven Rechtsdurchsetzung abhalten. Zu denken ist etwa an mitunter hohe Verfahrenskosten sowie die Schwierigkeit, die Aussicht auf Erfolg innerhalb des Rechtswegs selbst adäquat einschätzen zu können.[92] Eine Hilfe bietet hierbei die Künstliche Intelligenz, die bereits heute in einer Vielzahl von Anwendungen einen Zugang zum Recht gewährt.[93] Dabei sind prinzipiell zwei Einsatzformen von KI denkbar: eine – jeweils im Hinblick auf die menschliche Rechtsberatung – unterstützende sowie eine ersetzende.[94] Beispiele für den unterstützenden Einsatz von KI liefern Online-Informationssysteme für Rechtssuchende oder Aktenorganisationsprogramme.[95] Letztere verringern den Zeitbedarf bei der Bearbeitung einzelner Fälle und können damit Rechtsberatung günstiger machen, weshalb sie verbreitet befürwortet werden.[96] Allerdings können auch lediglich unterstützende KI-Systeme rechtliche Schwierigkeiten aufwerfen, sofern sie die Entscheidung der Rechtsanwenderin oder des Rechtsanwenders beeinflussen. Zu denken ist etwa an Verfahren der „predictive justice", die den Ausgang von Gerichtsentscheidungen prognostizieren. Werden sie durch Richterinnen und Richter eingesetzt, kann dies ihre Entscheidungsfreiheit beeinträchtigen,[97]

---

[91] Zum Risiko des „Kompetenzverlusts" im Medizinkontext *Hermann/Stock*, in: Interdisziplinäre Arbeitsgruppe „Verantwortung: Maschinelles Lernen und Künstliche Intelligenz" der Berlin-Brandenburgischen Akademie der Wissenschaften (Hrsg.), Kompetent eigene Entscheidungen treffen? Auch mit Künstlicher Intelligenz!, 2020, 24 (32).

[92] Vgl. *Tito*, How AI can improve access to justice, 23.10.2017, https://www.centreforpublicimpact.org/insights/joel-tito-ai-justice (letzter Zugriff 6.11.2021); zum Thema „Rechtskenntnis" siehe auch *Baer*, Rechtssoziologie, 4. Aufl., 2021, § 7 Rn. 10 f.

[93] Für Beispiele siehe *Wu*, AI Goes to Court: The Growing Landscape of AI for Access to Justice, 5.8.2019, https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to- justice-3f58aca4306f (letzter Zugriff 6.11.2021).

[94] *Tito* (Fn. 91).

[95] *Tito* (Fn. 91).

[96] *Tito* (Fn. 91).

[97] Vgl. *Rühl*, in: Kaulartz/Braegelmann (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 14.1 Rn. 29.

was den Einzelfall und die Rechte der Beteiligten aus dem Blick geraten lassen kann. Als ebenso problematisch haben sich Verfahren der Unterstützung im Strafrecht, beispielsweise bei Bewährungsentscheidungen erwiesen, sofern sie wie das US-amerikanische System COMPAS auf einem Bias beruhen.[98]

Besonders kritisch sind aus rechtlicher Perspektive solche KI-Systeme, durch die menschliche Rechtsanwenderinnen und Rechtsanwender teilweise oder vollständig ersetzt werden. Sofern dies angesichts der Schwierigkeiten der Formalisierung der Rechtssprache überhaupt möglich ist,[99] stoßen entsprechende Bestrebungen auf erhebliche Bedenken, die beispielsweise die Transparenz der KI-Systeme betreffen.[100] Das Rechtssystem beruht auf dem Austausch und der Vermittlung von Gründen für spezifische Freiheitseinschränkungen.[101] Sofern ein KI-System die einschlägigen Transparenzerfordernisse nicht erfüllt, vielmehr eine Black Box darstellt, entspricht es gerade nicht den Anforderungen, die an die Rechtsanwendung zu stellen sind.[102]

IX. Rechtsdurchsetzung bei verbrauchergefährdenden KI-Systemen

Ist es infolge des Einsatzes eines KI-Systems zu Verletzungen der Rechte von Verbraucherinnen und Verbrauchern gekommen, bedarf es der effektiven Rechtsdurchsetzung. Ein Instrument können hierbei Verbandsklagen bieten, die bereits aus anderem Kontext[103] bekannt sind. Sie eignen sich für die Rechtsdurchsetzung in Bezug auf verbrauchergefährdende KI-Systeme in besonderer Weise, weil durch entsprechende Anwendungen in aller Regel eine größere Zahl an Verbraucherinnen und Verbrauchern betroffen ist. Das Institut der Verbandsklage hat zuletzt durch die Verbandsklagen-Richtlinie (EU) 2020/1828 eine erhebliche Stärkung erfahren.[104] Auch behördliche Zulassungsverfahren sind eine Form der Rechtsdurchsetzung von Verbraucherinnen und Verbrauchern. Anders als in aller Regel Verfahren der Zertifizierung beruhen diese nicht auf der Freiwilligkeit des Her-

---

[98] Siehe zu diesem Thema *Eisele/Böhm*, in: Beck/Kusche/Valerius (Hrsg.), Digitalisierung, Automatisierung, KI und Recht. Festgabe zum 10-jährigen Bestehen der Forschungsstelle RobotRecht, 2020, 519 (527 f.) sowie *Nink*, Justiz und Algorithmen. Über die Schwächen menschlicher Entscheidungsfindung und die Möglichkeiten neuer Technologien in der Rechtsprechung, 2021, 376 ff.

[99] Vgl. *Tito* (Fn. 91).

[100] *Rostalski*, in: Hoven/Kudlich (Hrsg.), Digitalisierung und Strafverfahren, 2020, S. 263 (272 f.); *Rühl* (Fn. 96), Kap. 14.1 Rn. 19 f.

[101] Vgl. dazu *Rostalski* (Fn. 99), S. 272 f.

[102] Vgl. *Rostalski* (Fn. 99), S. 272 f.

[103] Zur Verbandsklage im Umweltrecht *Bunge*, JuS 2020, 740; zur Bedeutung von dieser im „Energiesektor" *Knauff*, EnWZ 2021, 3 ff. Beachte außerdem *Klocke*, Rechtsschutz in kollektiven Strukturen. Die Verbandsklage im Verbraucher- und Arbeitsrecht, 2016.

[104] Zu dieser *Röthemeyer*, VuR 2021, 43, sowie *Augenhofer*, NJW 2021, 113.

stellers eines KI-Systems, sondern sind verpflichtend.[105] Bei erheblich eingriffsintensiven Anwendungen erweist sich dies als angemessen, um die Rechte und Interessen von Verbraucherinnen und Verbrauchern zu schützen. Zulassungsverfahren können außerdem durch weitere Instrumente der behördlichen Aufsicht flankiert werden, die den weiteren Betrieb des KI-Systems begleiten.

X. Fazit

KI-Systeme bergen für Verbraucherinnen und Verbraucher eine Vielzahl an Chancen, aber auch Risiken. Aus diesem Grund erweisen sich KI-Systeme auch als notwendiger Gegenstand von Regulierung. In vielen Bereichen des Rechts sind insoweit noch Fragen offen. Die Verbraucherrechtstage 2021 schaffen vor diesem Hintergrund ein Problembewusstsein in Bezug auf die Notwendigkeit künftiger Regulierung und liefern erste Vorschläge zu ihrer Umsetzung.

[105] *Heesen* et al., Zertifizierung von KI-Systemen. Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme, S. 8, abrufbar unter https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1_3_Whitepaper_Zertifizierung_KI_Systemen.pdf (letzter Zugriff 6.11.2021).

Einsatz von KI-Systemen in der Justiz

Giesela Rühl[1]

I. Einleitung

Wenn es einen Lebensbereich gibt, in dem Künstliche Intelligenz (KI)[2] bislang kaum eine Rolle spielt, dann ist es die Justiz. Tatsächlich finden Systeme, die gemeinhin als künstlich intelligent bezeichnet werden, bislang lediglich in der US-amerikanischen Strafjustiz Verwendung, und zwar zur Beurteilung der Rückfallwahrscheinlichkeit von Straftätern.[3] Im Bereich der Ziviljustiz, die sich der gerichtlichen Beilegung privater Streitigkeiten widmet, fristet der Einsatz entsprechender Systeme demgegenüber bislang ein Schattendasein.[4] Lediglich die in Peking und Hangzou ansässigen chinesischen Internetgerichte, die Streitigkeiten aus online abgeschlossenen Geschäften beilegen, greifen dem Vernehmen nach auf „AI Judges" zurück.[5] Und Estland hat angekündigt, Klagen bis zu einem Wert von € 7.000,00 bald in erster Instanz durch „Roboterrichter" entscheiden lassen.[6]

Der nachfolgende Beitrag nimmt diesen Befund zum Anlass, um sich Gedanken über den Einsatz Künstlicher Intelligenz in der (deutschen) Ziviljustiz zu

---

[1] Der nachfolgende Beitrag ist ein (leicht aktualisierter) Nachdruck von *Giesela Rühl*, KI in der gerichtlichen Streitbeilegung, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 617 ff. Die Autorin dankt den Herausgebern und dem Verlag C. H. Beck für die Genehmigung zum Nachdruck.

[2] Siehe zum Begriff der Künstlichen Intelligenz *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15 ff.

[3] Siehe dazu unten III.2.b).

[4] Siehe dazu die Übersicht der *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 16 ff.

[5] Siehe dazu *Susskind*, Online Courts and the Future of Justice, 2019, 171 f.; *Young*, China Has Unveiled an AI Judge that Will ‚Help' With Court Proceedings, 19. August 2019, abrufbar unter https://interestingengineering.com/china-has-unveiled-an-ai-judge-that-will-help-with-court-proceedings (zuletzt abgerufen am 10.12.2021); *MacFadden*, Can AI be More Efficient Than People in the Judicial System, 4. Januar 2020, abrufbar unter https://interestingengineering.com/can-ai-be-more-efficient-than-people-in-the-judicial-system (zuletzt abgerufen am 10.12.2021). Siehe für einen Überblick über die Bemühungen Chinas zur Modernisierung der Justiz Supreme People's Court, Chinese Courts and Internet Judiciary, 4. Dezember 2019, abrufbar unter https://drive.google.com/file/d/1T8i303Czq1GV3RAbJc7tHXpSPxT2nv-5/view (zuletzt abgerufen am 10.12.2021);

[6] Siehe dazu Can AI Be a Fair Judge in Court? Estonia Thinks so, Wired, 5. März 2019, abrufbar unter https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ (zuletzt abgerufen am 10.12.2021).

machen. Dazu werden denkbare Einsatzmöglichkeiten vorgestellt (II.) sowie technische, systemische und rechtliche Herausforderungen aufgezeigt (III.). Aus Platzgründen beschränken sich die Überlegungen auf Systeme, die die in Zivilsachen tätigen ordentlichen Gerichte bei ihrer Tätigkeit unterstützen können. Überlegungen dazu, wie Künstliche Intelligenz Anwälten oder rechtssuchenden Bürger im Rahmen zivilgerichtlicher Verfahren helfen kann, bleiben demgegenüber außen vor.[7]

II. Einsatzmöglichkeiten

Der Begriff der Künstlichen Intelligenz ist schillernd und bislang weder in der Informatik noch in der Rechtswissenschaft einer abschließenden Definition zugeführt worden.[8] Nach dem derzeitigen Stand der Dinge umfasst er eine ganze Reihe ganz unterschiedlicher technischer Ansätze und Methoden, die alle in der einen oder anderen Weise versuchen, die geistige Leistung von Menschen nachzubilden, zu simulieren oder sogar zu übertreffen.[9] Im folgenden Kapitel wird der Begriff der Künstlichen Intelligenz deshalb nicht auf bestimmte technische Ansätze und Methoden beschränkt, zumal diese in der Praxis häufig kombiniert werden.[10] Vielmehr sollen aus der Menge der Systeme, die heute mit dem Begriff der Künstlichen Intelligenz in Verbindung gebracht werden, funktional solche in den Blick genommen, die sich perspektivisch für die Justiz fruchtbar machen lassen. Diese lassen sich in drei verschiedene Gruppen einteilen.

1. Dokumentenanalyse

Der ersten Gruppe sind Systeme zuzuordnen, die bei der Suche nach Dokumenten und Texten mit einem bestimmten Inhalt helfen und sich dabei Technologien des maschinellen Lernens[11] zu Nutze machen *(document review, document analysis, e-discovery)*.[12] Ihr Einsatz ist bereits heute weit verbreitet, und zwar insbeson-

---

[7] Siehe dazu *Fries*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 651 ff.

[8] Siehe dazu ausführlich *Herberger*, NJW 2018, 2825 ff. sowie *Braegelmann/Kaulartz*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 1 ff.

[9] *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15. Ebenso *Susskind*, Online Courts and the Future of Justice, 2019, 264 f.

[10] Siehe dazu, *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15 (16 ff.).

[11] Siehe zum Begriff des maschinellen Lernens *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15 (18 ff.).

[12] Siehe dazu für den deutschen Rechtsmarkt http://www.tobschall.de/legaltech unter „AI/eDiscovery/Automation Tools" sowie https://www.legal-tech-in-deutschland.de unter „Dokumentenanalyse". Siehe außerdem für den globalen Rechtsmarkt https://techindex.law.stanford.edu unter „E-Discovery".

dere dort, wo es um die Sichtung und Analyse großer Dokumentenmengen geht. Als Beispiel für ein System, dass dieser Gruppe zuzuordnen ist, sei hier nur die Analyse-Software KIRA genannt.[13] Sie kann innerhalb kürzester Zeit bestimmte Vertragsklauseln (Change-of-Control Klauseln, Wettbewerbsverbote, Haftungsbegrenzungen, etc.) in Dokumenten finden und hilft mittlerweile einer großen Zahl internationaler Kanzleien[14] bei der Bewertung rechtlicher Risiken im Rahmen von M&A-Transaktionen.[15] In der Justiz könnten entsprechende Systeme Richtern perspektivisch dabei helfen, einschlägige Urteile schneller und vor allen Dingen zuverlässiger zu finden. So ließe sich zum Beispiel daran denken, im Rahmen der gerichtlichen AGB-Kontrolle Software zum Einsatz zu bringen, die prüfen kann, ob eine Klausel bereits von anderen Gerichten für unwirksam erklärt wurde.

2. Dokumentenerstellung

Die zweite Gruppe von Systemen, die perspektivisch die Arbeit der Justiz erleichtern könnte, unterstützt die Erstellung standardisierbarer Dokumente (*document automation, document generation*).[16] Auch sie werden heute bereits vielfach in der Praxis, namentlich von Anwaltskanzleien eingesetzt. Zudem können Verbraucher und Unternehmer über verschiedene Online-Plattformen rechtlich relevante Dokumente (Verträge, Testamente, etc.) selbst generieren. Prominentes Beispiel für eine derartige Plattform ist der Dokumentengenerator Smartlaw von Wolters Kluwer.[17] Er gewährt Zugriff auf mehr als 190 verschiedene Rechtsdokumente, die mit Hilfe eines nutzerfreundlichen Frage-Antwort-Systems in kürzester Zeit „in Anwaltsqualität" auf die individuellen Bedürfnisse des Nutzers zugeschnitten werden können. In der Justiz könnten sich Richter perspektivisch vergleichbare Systeme zu Nutze machen, um Urteile oder Teile von Urteilen – oder zumindest Urteilsentwürfe – schneller zu erstellen.

---

[13] https://kirasystems.com. Ähnliche Angebote finden sich bei leverton.ai, rfrnz.com, thingsthinking.net.

[14] Siehe dazu die Nachweise unter https://kirasystems.com.

[15] Siehe dazu ausführlich *Krause/Hecker*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 83 ff.

[16] Siehe dazu für den deutschen Rechtsmarkt http://www.tobschall.de/legaltech unter „Contract Assembly & Tools" sowie https://www.legal-tech-in-deutschland.de unter „Dokumentenerstellung". Siehe https://techindex.law.stanford.edu (unter „Document automation").

[17] https://www.smartlaw.de. Ähnliche Angebote finden sich auf den Seiten von formblitz.de, hdcm.de, janolaw.de, legalos.io, wonder.legal, 123recht.de, foundersbox.vc, lawlift.de oder synergist.io. Im angloamerikanischen Rechtsraum erfreuen sich beispielsweise die Angebote von LegalZoom.com, ContractExpress.com, Hotdocs.com, Rocketlawyer.com, LawDepot.com, Nolo.com oder eForms.com großer Beliebtheit.

3. Entscheidungsvorhersage

Zur dritten Gruppe von Systemen, die sich perspektivisch für die Justiz fruchtbar machen lassen dürften, gehören schließlich Systeme, die sich mit der statistischen Auswertung von Urteilen und der datengestützten Vorhersage von Entscheidungen befassen (*outcome prediction, predictive analytics*).[18] Anders als die Systeme, die den ersten beiden Gruppen zuzuordnen sind, ist die Anzahl von Systemen, die sich der Entscheidungsvorhersage in der Justiz widmen bislang überschaubar groß. Sie sind aber der Nährboden für Überlegungen zum Einsatz von „Roboterrichtern", die den menschlichen Richter im Kernbereich seiner Tätigkeit, nämlich bei der Entscheidungsfindung unterstützen oder – in den kühnsten Träumen – ersetzen soll.[19] Das für Juristen erstaunliche – und für viele beängstigende – dabei ist: Anders als klassische, explizit programmierte Expertensysteme, die versuchen, die für die Entscheidung eines Falles erforderlichen Rechtsregeln in komplexen Entscheidungsbäumen abzubilden, und Ergebnisse deduktiv ableiten,[20] wenden die einschlägigen Systeme keine Rechtsregeln an, um ihre Vorhersagen zu treffen. Vielmehr entwickeln sie ihre Ergebnisse induktiv aus einer großen Menge an historischen Daten *über* Entscheidungen oder Fälle, die Muster und Zusammenhänge erkennen lassen, die für den Ausgang von Rechtsstreiten zumindest statistisch von Bedeutung sind. In Abhängigkeit davon, welche Daten für die Vorhersagen nutzbar gemacht werden, lassen sich zwei verschiedene Arten von Systemen unterscheiden.

a) Metadaten-Analyse

Die erste Art von Systemen greift für die Vorhersage auf sogenannte Metadaten zurück. Als Pionier kann hier das an der University of Stanford entwickelte System LexMachina gelten.[21] Es sagt die Wahrscheinlichkeit, eine Patentstreitigkeit in den USA zu gewinnen oder zu verlieren – angeblich – präziser voraus als US-amerikanische Patentanwälte – und das ohne jede Kenntnis des US-amerikanischen Patentrechts. Vorhersagegrundlage sind vielmehr Informationen über mehr als 100.000 Patentrechtsfälle, darunter die Namen der zuständigen Richter, der betei-

---

[18] Siehe dazu *Bues*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 275 (280); *Bull/Steffek*, ZKM 2018, 165 ff.; *Scherer*, J. Int'l Arb. 36 (2019) 539 (546 ff.); *Vogl*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 53 ff. sowie *Fries*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 651 (653 ff.) Siehe kritisch zum Begriff „predictive analytics" *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 57.

[19] Siehe dazu etwa *Fries*, NJW 2016, 2860 (2864); *Fries*, LTO v. 9. März 2018.

[20] Siehe dazu *Susskind*, Online Courts and the Future of Justice, 2019, 266 ff. sowie *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15 (22).

[21] https://lexmachina.com/. Siehe dazu *Susskind*, Online Courts and the Future of Justice, 2019, 282 f.

ligten Anwaltskanzleien und Rechtsanwälte sowie die Natur und der Wert des Streitgegenstandes. In ähnlicher Weise arbeiten die Produkte von LexPredict[22] (jetzt Elevate Services), Ravel Law[23] oder Predictice.[24]

Auf die Auswertung von Metadaten setzen außerdem Systeme, die in den letzten Jahren in wissenschaftlichen Studien verwendet wurden, um Entscheidungen des US-amerikanischen Supreme Court vorherzusagen.[25] In einer ersten Studie ließen Forscher im Jahr 2002 einen Computer sowie eine Gruppe von 83 Experten vorhersagen, ob der Supreme Court ein vorinstanzliches Urteil bestätigen oder aufheben würde.[26] Anders als die Experten, die auf alle Informationen einschließlich des anwendbaren Rechts zurückgreifen durften, legte der Computer seiner Vorhersage lediglich sechs Faktoren zugrunde, nämlich 1) den Gerichtsbezirk der Vorinstanz, 2) das Rechtsgebiet, 3) die Identität (Kategorie) des Klägers, 4) die Identität (Kategorie) des Beklagten, 5) die ideologische Ausrichtung der vorinstanzlichen Entscheidung (liberal oder konservativ) und 6) die Grundlage der Klage (Verfassungswidrigkeit eines Gesetzes oder nicht).[27] Als Trainingsdaten dienten 628 Entscheidungen des Supreme Court aus der Zeit von 1994 bis 2002.[28] Das Ergebnis: Der Computer lag in 75 % der Fälle richtig, die menschlichen Experten nur in 59,1 %.[29] In einer zweiten Studie wurde dieses Ergebnis im Jahr 2017 im Wesentlichen bestätigt.[30] Dieses Mal musste der Computer mit Hilfe einer deutlich größeren Zahl von Metadaten[31] 28.000 historische Entscheidungen des Gerichts aus der Zeit von 1816 bis 2015 vorhersagen, und zwar nur anhand von Informationen, die vor der Entscheidung zur Verfügung standen und die nichts mit dem anwendbaren Recht zu tun hatten. Erneut lag der Computer in über 70 % der Fälle richtig.[32]

---

[22] https://www.lexpredict.com.

[23] https://home.ravellaw.com.

[24] https://predictice.com.

[25] Siehe dazu ausführlich *Bull/Steffek*, ZKM 2018, 165 (166 ff.); *Scherer*, J. Int'l Arb. 36 (2019) 539 (547 ff.). Andere Studien widmen sich beispielsweise der Vorhersage von Entscheidungen nach dem US-amerikanischen Bankruptcy Act. Siehe dazu *Warren*, Ann. Surv. of Bankr. Law 13 (2018 WL 4293106), abrufbar unter https://ssrn.com/abstract=3183484 oder http://dx.doi.org/10.2139/ssrn.3183484 (zuletzt abgerufen am 10.12.2021).

[26] *Ruger/Kim/Martin/Quinn*, Colum. L. Rev. 104 (2004) 1150 (1160 ff.). Siehe dazu auch *Bull/Steffek*, ZKM 2018, 165 (166).

[27] *Ruger/Kim/Martin/Quinn*, Colum. L. Rev. 104 (2004) 1150 (1163 f).

[28] *Ruger/Kim/Martin/Quinn*, Colum. L. Rev. 104 (2004) 1150 (1163).

[29] *Ruger/Kim/Martin/Quinn*, Colum. L. Rev. 104 (2004) 1150 (1171 ff.).

[30] *Katz/Bommarito II/Blackman*, PloS One 12:e0174698 (2017) 1 ff. Siehe dazu *Bull/Steffek*, ZKM 2018, 165 (166 f.); *Scherer*, J. Int'l Arb. 36 (2019) 539 (550 ff.); *Vogl*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 53 (56).

[31] *Katz/Bommarito II/Blackman*, PloS One 12:e0174698 (2017) 1 (4 ff.).

[32] *Katz/Bommarito II/Blackman*, PloS One 12:e0174698 (2017) 1 (8).

b) Sachverhaltsanalyse

Die zweite Art von Systemen ist – bislang – weniger verbreitet als die erste Art. Sie stützt ihre Vorhersagen nicht auf Metadaten, sondern auf die konkret zu entscheidenden Sachverhalte und vergleicht sie mit einschlägigen Vorentscheidungen. Sie wurde beispielsweise in einer Studie zur Vorhersage von Entscheidungen des Europäischen Gerichtshofs für Menschenrechte (EGMR) gesetzt[33]. Die beteiligten Wissenschaftler fütterten hier die Künstliche Intelligenz zunächst in Textform mit den Sachverhalten von knapp 600 Entscheidungen zu Art. 3 EMRK (Verbot der Folter), Art. 6 EMRK (Recht auf ein faires Verfahren) und Art. 8 EMRK (Recht auf Achtung des Privat- und Familienlebens).[34] Sodann wurde sie mit 10 % dieser Daten trainiert. Für die restlichen 90 % musste das System schließlich vorhersagen, ob der EGMR eine Verletzung der jeweiligen Vorschriften angenommen hatte ohne nicht. In erstaunlichen 79 % der Fälle lag es richtig.[35]

Noch bessere Ergebnisse erzielte ein ähnliches System in einer Studie zur Vorhersage der Entscheidungen der englischen Ombudsstelle für Finanzdienstleistungen.[36] Die Künstliche Intelligenz Case Cruncher Alpha musste hier bestimmen, ob ein bestimmtes Kreditversicherungsprodukt rechtmäßig verkauft worden war. Dazu wurde es mit 100.000 vergangenen Entscheidungen der Ombudsstelle gefüttert, in der genau diese Frage entschieden worden war. Mit einer Quote von 86,6 % sagte Case Cruncher Alpha daraufhin zukünftige Entscheidungen der Ombudsstelle voraus. Eine Vergleichsgruppe von etwa 100 Wirtschaftsanwälten kam lediglich auf 62,3 %.

III. Herausforderungen

Es bedarf keiner besonderen Erläuterung, dass der Einsatz von KI-Systemen der gerade beschriebenen Art mit enormen Chancen für die Justiz verbunden wäre.[37] Denn wenn der gesamte Rechtsprechungsbestand zu einer bestimmten Rechts-

---

[33] *Aletras/Tsarapatsanis/Preotiuc-Pietro/Lampos*, PeerJ Computer Science 2:e93 (2016) 1 ff. Siehe dazu *Bull/Steffek*, ZKM 2018, 165 (166); *Scherer*, J. Int'l Arb. 36 (2019) 539 (547 ff.).

[34] Art. 3, 6 und 8 EMRK wurden ausgewählt, weil zu diesen Vorschriften die meisten Entscheidungen vorlagen. *Aletras/Tsarapatsanis/Preotiuc-Pietro/Lampos*, PeerJ Computer Science 2:e93 (2016) 1 (6).

[35] *Aletras/Tsarapatsanis/Preotiuc-Pietro/Lampos*, PeerJ Computer Science 2:e93 (2016) 1 (10).

[36] https://www.case-crunch.com/#challenge. Siehe dazu *Steffek*, ZKM 2018, 75 (Editorial); *Bull/Steffek*, ZKM 2018, 165 (166 f.).

[37] *Fries*, NJW 2016, 2860 (2864); *Fries*, LTO v. 9. März 2018; *Hanke*, Transnat'l Disp. Mgmt. 14 (2017–2) 1 (5 ff.); *Steffek*, ZKM 2018, 75 (Editorial); *Susskind*, Online Courts and the Future of Justice (2019) 277 ff. Skeptischer demgegenüber *Scherer*, J. Int'l Arb. 36 (2019) 53 (554 ff., 572 f.); *Scherer*, Austrian Y.B. Int'l Arb. 2019, 503 (509 ff.) und die *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 213 und 218, die den Einsatz „algorithmischer Systeme" in der Rechtsprechung für „hochproblematisch" und allenfalls in den „Randbereichen" für zulässig

frage in Sekundenschnelle ausgewertet, Entscheidungsvorschläge gemacht und Urteilsentwürfe automatisiert erstellt werden könnten, dann würde dies zu einer Entlastung der Justiz und einer Beschleunigung von Verfahren führen. Zudem könnten Fehler vermieden, die Einheit der Rechtsordnung sichergestellt sowie die Gleichheit vor Gericht garantiert werden. Insbesondere könnte vermieden werden, dass außerrechtliche Faktoren Einfluss auf Entscheidungen nehmen, die keinen Einfluss auf Entscheidungen nehmen sollten.[38] Allerdings: Was in der Theorie verlockend klingt, begegnet in der Praxis einer ganzen Reihe von Herausforderungen.

1. Technische Herausforderungen

Die erste Herausforderung ist – wenig überraschend – technischer Art:[39] Künstliche Intelligenz, die die richterliche Tätigkeit in ihrer gesamten Breite oder auch nur in Teilen übernehmen oder richterliche Tätigkeit sinnvoll ergänzen könnten, gibt es bislang noch nicht. Die oben beschriebenen Systeme sind auf bestimmte Rechtsgebiete, bestimmte Fälle und bestimmte Fragen beschränkt. So kann KIRA nur bestimmte Vertragsklauseln erkennen, Smartlaw nur bestimmte Rechtsdokumente erstellen und Case Cruncher Alpha nur eine bestimmte Rechtsfrage beantworten.

Freilich: Die Entwicklung von KI-Systemen für die Justiz befindet sich erst am Anfang und niemand weiß, wo wir in 10, 20 oder 30 Jahren stehen werden. Sagen lässt sich allerdings, dass die Entwicklung einsatzfähiger Systeme nicht so leicht sein wird, wie viele meinen.[40] So verlangt schon die Beantwortung kleinerer Rechtsfragen die Programmierung äußerst komplexer und umfassender Expertensysteme.[41] Und für den Einsatz selbstlernender Systeme müssen Urteile 1) in großer Zahl und 2) in strukturierter und maschinenlesbarer Form vorliegen.[42] Beides ist jedoch, zumindest derzeit und zumindest in Deutschland nicht der

---

hält. Differenzierend *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 41 ff.

[38] Siehe dazu nur die mittlerweile legendäre israelische Studie von *Danziger/Levav/Avnaim-Pesso*, PNAS 108 (2011) 6889, abrufbar unter https://www.pnas.org/content/108/17/6889 (zuletzt abgerufen am 10.12.2021), die zeigt, dass die Wahrscheinlichkeit, dass ein Straftäter auf Bewährung entlassen wird, vor der Mittagspause des Richters deutlich geringer ist, als danach.

[39] Siehe dazu auch *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15 (30 f.).

[40] *Susskind*, Online Courts and the Future of Justice, 2019, 263. Ebenso *Adrian*, Rechtstheorie 48 (2017) 77 (121).

[41] Siehe dazu *Susskind*, Online Courts and the Future of Justice, 2019, 266 ff. unter Hinweis auf das von ihm selbst in den 1980er Jahren programmierte System zur Beantwortung der Frage nach der Verjährung bestimmter Ansprüche, das am Ende aus einem Entscheidungsbaum mit mehr als 2 Millionen Pfaden bestand.

[42] Siehe dazu *Scherer*, J. Int'l Arb. 36 (2019) 555 (554 f.); *Vogl*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 53 (60); *von Bünau*, in: Breidenbach/Glatz (Hrsg.) Rechtshandbuch Legal Tech, 2021, 71 (78). Siehe außerdem *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 18 f.

Fall.[43] Tatsächlich wird nur ein Bruchteil aller Urteile, die jedes Jahr von deutschen Gerichten erlassen werden, in juristischen Fachzeitschriften und Datenbanken veröffentlicht oder über das Online-Rechtsprechungsportal von BMJV und Bundesamtes für Justiz[44] frei zur Verfügung gestellt.[45] Die allermeisten Urteile verschwinden demgegenüber in den Schubladen der Justiz und erblicken nie das Licht der Öffentlichkeit.[46] In der Regel werden sie sogar nach einer gewissen Zeit vernichtet, so dass sie auch nachträglich nicht mehr zugänglich gemacht werden können.[47] Dies gilt insbesondere für die große Masse an erstinstanzlichen, amts- und landgerichtlichen Entscheidungen, die rechtlich nichts Neues bringen, aber für die Nutzung selbstlernender System von entscheidender Bedeutung wären. Bei den in juristischen Fachzeitschriften veröffentlichten Entscheidungen kommt hinzu, dass sie urheberrechtlich geschützt und deshalb von externen Anbietern nicht ohne weiteres genutzt und ausgewertet werden dürfen.[48] Dem Einsatz selbstlernender Systeme steht schließlich auch entgegen, dass selbst veröffentlichte Urteile in der Regel nur in Textform vorliegen, die für KI-Systeme schwer zugänglich ist, weil die Extraktion und Verarbeitung von Informationen aus Texten sowie ihre Überführung in maschinenlesbare Datensätze (*natural language processing*) noch in den Kinderschuhen steckt.[49] Bis es tatsächlich Systeme gibt, die in der Justiz eingesetzt werden können, müssen deshalb noch viele technische Hürden überwunden werden.

2. Systemische Herausforderungen

Neben den technischen Herausforderungen treffen KI-Systeme aber auch auf systemische Herausforderungen. Drei von ihnen sollen im Folgenden kurz beschrieben werden. Sie beziehen sich vornehmlich auf Systeme, die ihre Ergebnisse nicht

---

[43] Besser sieht es in anderen Ländern, namentlich in Frankreich und den USA aus. Siehe dazu *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 20 ff.

[44] http://www.rechtsprechung-im-internet.de.

[45] *Coupette/Fleckner*, JZ 2018, 379 (380 ff.); *Braegelmann*, Lack of Data, Law of Law, 14. Februar 2019, abrufbar unter https://www.linkedin.com/pulse/lack-data-law-tom-braegelmann/ (zuletzt abgerufen am 10.12.2021); *Fries*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 651, 655 f.

[46] *Coupette/Fleckner*, JZ 2018, 379 (380 f.); *Braegelmann*, Lack of Data, Law of Law, 14. Februar 2019, abrufbar unter https://www.linkedin.com/pulse/lack-data-law-tom-braegelmann/ (zuletzt abgerufen am 10.12.2021); *Fries*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 651 (655 f.).

[47] Siehe zu diesem Problem *Coupette/Fleckner*, JZ 2018, 379 (381 f.).

[48] So LG Berlin, Urt. v. 5. Mai 2015, BeckRS 2015, 15707.

[49] *von Bünau*, in: Breidenbach/Glatz (Hrsg.) Rechtshandbuch Legal Tech, 2021, 71 (78 ff.); *Bues*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 275 (284 f.). Siehe dazu aber das Forschungsprojekt von *Bull/Steffek*, in: Aggarwal/Eidenmüller/Enriques/Payne/van Zwieten (Hrsg.), Autonomous Systems and the Law, 2019, 67 ff., das sich um die Aufbereitung von 100.000 US-amerikanischen Fällen bemüht.

deduktiv und regelbasiert, sondern induktiv durch Auswertung und Analyse großer Mengen an historischen Daten, namentlich durch maschinelles Lernen erzielen.

a) Versteinerungsgefahr

Die erste Herausforderung wird in der Literatur typischerweise als Versteinerungsgefahr (oder auch: Konservatismus) bezeichnet.[50] Sie findet ihren Ursprung in dem Umstand, dass datenbasiert arbeitende KI-Systeme ihre Erkenntnisse aus historischen Datensätzen – und damit aus vergangenen Ereignissen – ableiten. Ihnen wird deshalb vorgeworfen, mit neuen, unbekannten Fragen und Problemen nicht umgehen und außerdem keine neuen, innovativen Lösungen entwickeln zu können.[51] Für den Einsatz in der Justiz würde dies bedeuten, dass KI-Systeme allenfalls zur Entscheidung von Rechtsfragen eingesetzt werden könnten, die in der Vergangenheit bereits entschieden wurden.

Bei genauerer Betrachtung ist das Problem der Versteinerung freilich kleiner als man denkt.[52] Denn auch der menschliche Richter kann nicht in die Zukunft blicken. Auch der menschliche Richter löst neue Fälle nur unter Rückgriff auf historische Daten. Auch der menschliche Richter hat nicht mehr als die Vergangenheit als Referenzpunkt, um sich mit neuen Problemen auseinanderzusetzen. Anders als künstlich intelligente Systeme greifen sie allerdings nie auf alle, sondern immer nur auf einen kleinen Teil der zur Verfügung stehenden Daten zurück. Ein selbstlernendes KI-System kann sich demgegenüber in kürzester Zeit Zugriff zu allen Daten – beispielsweise zum gesamten Rechtsprechungsbestand – beschaffen und auswerten.[53] Hinzukommt, dass der Datenbestand, auf den selbstlernende Systeme zugreifen, ständig aktualisiert werden kann – und das in einer Perfektion, die kein Mensch für sich in Anspruch nehmen kann. KI-Systeme können deshalb durchaus Ergebnisse liefern, die überraschend sind und – wenn sie von Menschen erzeugt würden – als innovativ oder kreativ bezeichnet werden könnten.[54] Aber selbst wenn dies nicht der Fall ist, wird man sagen können, dass KI-Systeme zumindest bei der Behandlung standardisierbarer Fälle, die keine neuen Probleme aufwerfen und weniger Innovationskraft und Kreativität verlangen als Schnelligkeit, perspektivisch gute Dienste leisten können.[55] Auch bei diesen bleibt freilich

---

[50] Siehe dazu *Scherer*, J. Int'l Arb. 36 (2019) 555 (557).

[51] *Scherer*, J. Int'l Arb. 36 (2019) 555 (557). Siehe aber *Susskind*, Online Courts and the Future of Justice, 2019, 277 ff.

[52] *Susskind*, Online Courts and the Future of Justice, 2019, 289 f.

[53] *Susskind*, Online Courts and the Future of Justice, 2019, 289 f.

[54] *Susskind*, Online Courts and the Future of Justice, 2019, 270 und 289, unter Hinweis auf den – vielfach als innovative und kreativ beschriebenen – 37. Zug des 2. Spiels zwischen der von Google Deep Mind geschaffenen künstlichen Intelligenz AlphaGo und dem weltbesten Go-Spieler Lee Sedol.

[55] *Susskind*, Online Courts and the Future of Justice, 2019, 290.

ein Problem: Ändern sich gesetzliche Vorgaben oder die – in Deutschland zwar nicht bindende, aber trotzdem autoritative – höchstrichterliche Rechtsprechung, führt der Rückgriff auf die Auswertung von Urteilen, die vor der Änderung ergangen sind, zu Fehlern.[56] Wie sichergestellt werden kann, dass nicht mehr relevante Urteile ignoriert oder nur soweit weiterverwendet werden, als sie von Änderungen nicht betroffen sind, ist bislang unklar.

b) Diskriminierungsgefahr

Ein weiteres – mittlerweile weithin bekanntes und weithin diskutiertes – systemisches Problem von KI-Systemen ist daneben ihr Potential, Personen zu diskriminieren.[57] Als paradigmatisches Beispiel gilt insofern das COMPAS System (Correctional Offender Management Profiling for Alternative Sanctions), das in zahlreichen Bundesstaaten der USA eingesetzt wird, um die Rückfallwahrscheinlichkeit von Straftätern zu bestimmen.[58] Es berücksichtigt die Antworten der Delinquenten auf 137 Fragen – und stuft danach schwarze Männer mehr als doppelt so häufig fälschlicherweise als rückfallgefährdet ein wie weiße Männer, obwohl nach der Hautfarbe nicht direkt gefragt wird. Hinweise auf Diskriminierungen gibt es aber auch bei anderen Systemen, zum Beispiel solchen, die bei der Auswahl von Arbeitnehmern, bei der Vermietung von Wohnungen, beim Abschluss von Verträgen im elektronischen Geschäftsverkehr oder bei der Vergabe von Krediten eingesetzt werden.[59] Die Ursachen für Diskriminierung lassen sich dabei nicht immer genau feststellen. Feststeht jedoch, dass sie vielfältig und komplex sind.[60] Neben gezielter Manipulation oder etwaigen Vorurteilen des Ent-

---

[56] *Scherer*, J. Int'l Arb. 36 (2019) 555 (557). Ebenso im Ergebnis *von Bünau*, in: Breidenbach/Glatz (Hrsg.) Rechtshandbuch Legal Tech, 2021, 71 (81 ff.); *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 66.

[57] Siehe dazu ausführlich und mit zahlreichen Beispielen *Barocas/Selbst*, Calif. L. Rev. 104 (2016) 671 ff.; *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019; *Zuiderveen Borgesius*, Discrimination, artificial intelligence and algorithmic decision-making, 2018.

[58] *Angwin/Larson/Mattu/Kirchner*, ProPublica, 23. Mai 2016, abrufbar unter, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; *Larson/Mattu/Kirchner/Angwin*, How We Analyzed the COMPAS Recidivism Algorithm, ProPublica, 23. Mai 2016, abrufbar unter https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (zuletzt abgerufen am 10.12.2021) sowie *Zuiderveen Borgesius* (Fn. 57), 14 ff. Siehe außerdem den Überblick von *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, 66 f., *Zuiderveen Borgesius* (Fn. 57), 14 f. sowie der *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 48 ff.

[59] Siehe dazu ausführlich *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, 34 ff., 41 ff., 44 ff. und 49 ff.; *Zuiderveen Borgesius* (Fn. 57), 15 ff. Siehe außerdem den Überblick bei *Busch*, Algorithmic Accountability, 2018, 20 ff.

[60] Siehe dazu ausführlich *Barocas/Selbst*, Calif. L. Rev. 104 (2016) 671 (678 ff.); *Orwat*, Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, 76 ff.; *Zuiderveen Borgesius* (Fn. 57), 10 ff. Siehe außerdem den Überblick bei *Busch*, Algorithmic Accountability, 2018, 20 ff.

wicklers können beispielsweise die verwendeten historischen Daten vorbelastet sein und deshalb bestehende Diskriminierungen schlicht weitertragen. Im Kontext der Urteilsanalyse ließe sich insofern an Urteile aus der Zeit des 3. Reichs denken, in denen die – bis heute im Wortlaut unverändert geltenden Generalklauseln des BGB – durch Gerichte genutzt wurden, um die Benachteiligung von Juden zu rechtfertigen.

Für die Justiz ist dieser Befund selbstredend ein ernsthaftes Problem. Tatsächlich lässt sich wohl kaum ein Bereich vorstellen, in dem Neutralität und Gleichbehandlung wichtiger wären als vor staatlichen Gerichten. Die Diskriminierungsfreiheit der einschlägigen Systeme sicherzustellen, dürfte deshalb perspektivisch eine der größten Herausforderungen für den Einsatz künstlich intelligenter Systeme in der Justiz sein.[61] Wegen der Vielfältigkeit der Diskriminierungsursachen ist sie freilich auch besonders schwer zu bewältigen.

c) Black box-Problem

Ein letztes systemisches Problem, das der Einsatz Künstlicher Intelligenz in der Justiz mit sich bringt und hier erwähnt werden soll, ist schließlich das sogenannte black box-Problem.[62] Es beschreibt den Umstand, dass sowohl die Arbeitsweise als auch das Ergebnis der einschlägigen Systeme für den Anwender – und in der Regel auch für den Entwickler – nicht nachvollziehbar sind.[63] Das System arbeitet wie eine black box, die mit Daten gefüttert wird und am Ende ein Ergebnis auswirft, ohne dass erkennbar ist, was dazwischen geschieht. In der Justiz hätte der Einsatz Künstlicher Intelligenz deshalb zur Folge, dass Richter die erzielten Ergebnisse weder nachvollziehen noch prüfen könnten, wie das Ergebnis zustande gekommen ist. Sie könnten es dementsprechend auch nicht kritisch hinterfragen, geschweige denn erklären oder begründen. All dies ist aber gerade in der Justiz von essentieller Bedeutung[64] und deshalb auch als Ausfluss des Anspruchs auf rechtliches Gehör (Art. 103 Abs. 1 GG)[65] verfassungsrechtlich garantiert.[66] Denn

---

[61] Ebenso *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 9.

[62] *Bues*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech. 2018, 275 (283); *Scherer*, J. Int'l Arb. 36 (2019) 555 (562 ff.); *Vogl*, in: Hartung/Bues/Halbleib (Hrsg.), Legal Tech, 2018, 53 (60 f.). Siehe außerdem *Stiemerling*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 15 (26).

[63] Dazu *Körner*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 44 ff.

[64] *Braegelmann*, Lack of Data, Law of Law, 14. Februar 2019, abrufbar unter https://www.linkedin.com/pulse/lack-data-law-tom-braegelmann/ (zuletzt abgerufen am 10.12.2021); *Scherer*, J. Int'l Arb. 36 (2019) 555 (562 ff.). Ebenso die Einschätzung der *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 213 sowie – allgemein zum Gesetzesvollzug – *Meyer*, ZRP 2018, 233 (237).

[65] Siehe dazu auch noch unter III.2.b).

[66] Maunz/Dürig/*Remmert*, Grundgesetz, 88. EGL, Stand August 2019, Art. 103 Rn. 96.

ein Urteil ohne Erklärung und Begründung kann kaum zur Akzeptanz der Entscheidung und zur Schaffung von Rechtsfrieden beitragen. Wenn es anders wäre – und allein das Urteil zählen würde – könnten sich die Parteien auch darauf verständigen, ihren Streit durch Würfeln beizulegen. Dass dies in der Praxis nicht geschieht, deutet darauf hin, dass von einer gerichtlichen Entscheidung im Normalfall mehr – nämlich eine inhaltliche Auseinandersetzung sowie im Idealfall eine überzeugende Begründung – erwartet wird.[67] Hinzukommt, dass auch nur eine nachvollziehbare und begründete Entscheidung Verhaltensanreize für Dritte setzt, sich an die Entscheidung zu halten. Auch eine kritische Auseinandersetzung mit Gerichtsentscheidungen wird erst durch die Begründung ermöglicht.

KI-Systeme können vor diesem Hintergrund überhaupt nur dann für einen Einsatz in der Justiz in Betracht kommen, wenn sichergestellt ist, dass sie transparent arbeiten und die Ergebnisse so nachvollziehbar sind, dass sie vom Richter überprüft und zum Gegenstand einer Begründung gemacht werden können.[68] Hoffnungen, dass die Entwicklung entsprechender System gelingen kann, geben Forschungsansätze, die unter dem Schlagwort *explainable AI* in jüngster Zeit für Aufsehen gesorgt haben.[69] Sie zielen auf die Entwicklung von Mechanismen ab, die die Transparenz und die Nachvollziehbarkeit von KI-Systemen sicherstellen sollen, indem sie beispielsweise sogenannte *saliency maps* erstellen, die aufzeigen, welche Daten für das erzielte Ergebnis wichtig oder weniger wichtig waren.[70] Die Forschung hierzu steckt freilich noch in den Anfängen.

3. Rechtliche Herausforderungen

Schließlich trifft der Einsatz von KI-Systemen in der Justiz auch auf eine ganze Reihe von rechtlichen, insbesondere verfassungsrechtlichen Herausforderungen. Auf drei von ihnen soll hier kurz eingegangen werden.[71] Sie betreffen den Justizge-

---

[67] Damit ist nicht gesagt, dass es keine Fälle gibt, in denen die Parteien in erster Linie ein Interesse an einer irgendwie gearteten Entscheidung haben. *Susskind*, Online Courts and the Future of Justice, 2019, 290 verweist beispielsweise auf Länder, in denen die Gerichte einen so hohen Rückstand haben, dass eine Entscheidung im normalen Verfahren nicht zu erwarten ist. In Deutschland, wo ein erstinstanzliches Urteil im Schnitt innerhalb von 6 Monaten zu erlangen ist und zudem zahlreiche Stellen schnellen außergerichtlichen Rechtsschutz anbieten, dürften diese Fälle aber die Ausnahme bilden.

[68] Ebenso *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 12.

[69] Für einen Überblick siehe *Busch*, Algorithmic Accountability, 2018, 61 (m. w. N.); *The Royal Society*, Explainable AI: the basics, 2019; *Körner*, in: Braegelmann/Kaulartz (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, 44 ff.

[70] Siehe zu den verschiedenen Techniken *The Royal Society*, Explainable AI: the basics, 2019, 12 ff.

[71] Weitere Herausforderungen beziehen sich beispielsweise auf den Persönlichkeits- oder den Datenschutz. Siehe dazu nur *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 25 ff.

währungsanspruch (Art. 20 Abs. 3 i. V. m. Art. 2 GG), den Anspruch auf rechtliches Gehör (Art. 103 Abs. 1 GG) und die Unabhängigkeit des Richters (Art. 97 GG).

a) Justizgewährungsanspruch (Art. 20 Abs. 3 i. V. m. Art. 2 GG)

Der Justizgewährungsanspruch verpflichtet den Staat für die Beilegung privater Streitigkeiten eine Rechtsschutzmöglichkeit zur Verfügung zu stellen.[72] Im Bereich der Ziviljustiz kommt der Staat dieser Verpflichtung durch die Einrichtung von Zivilgerichten nach, die mit unabhängigen Richtern besetzt sind (Art. 92, 97 GG). Richter i. S. d. Grundgesetzes – und auch i. S. d. einfachen Rechts, namentlich des GVG und des DRiG – können aber zumindest *de lege lata* anerkanntermaßen nur Menschen sein.[73] Deshalb schließt das derzeit geltende Gerichtsverfassungsrecht eine vollständige Delegation richterlicher Tätigkeit auf künstlich intelligente Systeme – oder einen verpflichtenden Einsatz mit Übernahmeautomatismus – aus.[74] Angesichts der Tatsache, dass es in naher Zukunft keine Systeme geben wird, die den Richter vollständig oder auch nur in bestimmten Rechtsbereichen ersetzen könnten, ist diese Einschränkung freilich gut zu verkraften. Im Lichte der oben beschriebenen systemischen Herausforderungen von KI-Systemen, dürfte es sogar zu begrüßen ein, dass gerichtliche Entscheidungen bis auf Weiteres von einem Menschen getroffen und verantwortet werden müssen.

Möglich und zulässig dürfte es perspektivisch allerdings sein, gut geeignete Fälle – wie beim maschinellen Mahnverfahren (§ 689 Abs. 1 S. 2 ZPO)[75] – in einem ersten Schritt durch KI-Systeme entscheiden zu lassen und erst in einem zweiten Schritt auf Antrag der Parteien eine Überprüfung durch einen menschlichen Richter vorzunehmen.[76] Namentlich bei Standardfällen mit geringem Streitwert oder im einstweiligen Rechtsschutz, wo es auf Schnelligkeit ankommt und eventuelle Fehler im Hauptverfahren korrigiert werden können, dürften gute Argumente für den Einsatz künstlich intelligenter Systeme sprechen, zumindest wenn die oben beschriebenen technischen und systemischen Herausforderungen bewältigt werden können. So sieht man es wohl auch in Estland, wo die Pläne zur Einführung eines „Roboterrichters" zum einen auf Klagen bis zu einem Streitwert von € 7.000,00 beschränkt sind und der maschinellen Entscheidungen auf Antrag der Parteien zum anderen die Überprüfung durch einen menschlichen Richter nachfolgen soll.[77]

---

[72] BVerfGE NJW 2003, 1924; BeckOK GG/*Huster/Rux*, 41. Ed. 15.2.2019, Art. 20 GG Rn. 199.

[73] Ebenso *Enders*, JA 2018, 721 (723); *von Graevenitz*, ZRP 2018, 238 (240) sowie die *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 213.

[74] *Enders*, JA 2018, 721 (723); *von Graevenitz*, ZRP 2018, 238 (240).

[75] Siehe dazu *Sujecki*, MMR 2006, 369 (371 f.).

[76] Ebenso *von Graevenitz*, ZRP 2018, 238 (241).

[77] Can AI Be a Fair Judge in Court? Estonia Thinks so, Wired, 5. März 2019, abrufbar unter https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/.

Interessanter und ungleich schwerer zu beantworten ist demgegenüber die Frage, ob sich Richter im Rahmen ihrer Entscheidungsfindung perspektivisch durch Systeme Künstlicher Intelligenz unterstützen lassen dürfen. Bereits heute verlassen sich Richter auf die – algorithmengetriebenen – Suchmaschinen von juris und beck-online, um relevante Literatur und Rechtsprechung zu finden. Sollten sie, soweit vorhanden – noch bessere Systeme nutzen dürfen, um einschlägige Entscheidungen aufzufinden, auszuwerten und einen Hinweis zur Entscheidung des vorliegenden Falles zu erhalten? Aus Sicht des Justizgewährungsanspruchs lässt sich dagegen grundsätzlich nichts einwenden – solange am Ende ein menschlicher Richter die Entscheidung trifft und verantwortet. Zum Problem wird die Nutzung Künstlicher Intelligenz allerdings dann, wenn sie Richter dazu verleitet, sich ohne kritische Prüfung an den Ergebnissen des Computers zu orientieren und den Entscheidungsvorschlägen blind zu folgen. Dass diese Gefahr besteht, lässt sich nicht von der Hand weisen. Denn blindes Vertrauen in computergenerierte Entscheidungen ist weit verbreitet und hat unter dem Schlagwort *automation bias* Eingang in die wissenschaftliche Literatur gefunden.[78] Muss die Nutzung künstlich intelligenter Systeme also verboten werden, um der Gefahr zu begegnen, dass die Entscheidung des menschlichen Richters – *de facto* – durch die Entscheidung eines Computers ersetzt und der Justizgewährungsanspruch des Bürgers verletzt wird?[79]

Diese Schlussfolgerung zu ziehen, hieße aber wohl, das Kind mit dem Bade auszuschütten.[80] Denn zum einen würden Richter durch ein Nutzungsverbot langfristig – insbesondere im Vergleich zu Anwaltskanzleien – ins Hintertreffen geraten. Und zum anderen wäre ein Nutzungsverbot angesichts der vielen Vorteile, die mit dem Einsatz leistungsfähiger Systeme Künstlicher Intelligenz zumindest potentiell verbunden sind,[81] auch nicht wünschenswert. Dem *automation bias* sollte deshalb besser durch Ausbildung, Information und Aufklärung begegnet werden. Insbesondere müsste ein kritischer Umgang mit KI-Systemen vermittelt und die Leistungsfähigkeit und vor allen Dingen die Grenzen der eingesetzten Systeme aufgezeigt werden.[82] Schließlich dürfte auch kein Zweifel daran gelassen

---

[78] Ein fast schon legendäres Beispiel für *automation bias* ist der Fall dreier japanischer Studenten, die ihren Mietwagen im Vertrauen auf ihr GPS vor Australien ins Meer steuerten, um zu einer Insel zu gelangen. Siehe dazu *Hanson*, GPS Leads Japanese Tourists To Drive Into Australian Bay, The Huffington Post, 19. März 2012.

[79] So wohl die Einschätzung der *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 213 die den Einsatz von „algorithmischen Systemen" in der Justiz wegen des *automation bias* für „hochproblematisch" hält.

[80] Anders wohl die Einschätzung der *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 213 und 218, die den Einsatz „algorithmischer Systeme" in der Justiz auf „Randbereiche" beschränken möchte.

[81] Siehe dazu oben III.

[82] Ebenso *von Graevenitz*, ZRP 2018, 241 und *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), S. 12 sowie – ganz allgemein – die *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019), 166 f.

werden, dass letztendlich die handelnden Richter – und nicht das verwendete System – für die getroffenen Entscheidungen verantwortlich sind.[83]

b) Rechtliches Gehör (Art. 103 Abs. 1 GG)

Eine zweite rechtliche Herausforderung ergibt sich aus dem Anspruch auf rechtliches Gehör, der jedermann nach Art. 103 Abs. 1 GG vor staatlichen Gerichten zusteht. Er verlangt, dass tatsächliches und rechtliches Vorbringen von dem mit der Sache befassten Gericht – und damit von einem menschlichen Richter – zur Kenntnis genommen und bei der Entscheidung berücksichtigt wird.[84] Da die Rechtssuchenden darüber hinaus nicht zum bloßen Objekt des Verfahrens gemachten werden dürfen,[85] dürfte auch der Anspruch auf rechtliches Gehör eine vollständige Substituierung des menschlichen Richters durch eine Maschine ausschließen. Ein gestufter Einsatz Künstlicher Intelligenz, wie er in Estland in Planung ist und wie er beim maschinellen Mahnverfahren (§ 689 Abs. 1 S. 2 ZPO) auch in Deutschland bereits zur Anwendung kommt, dürfte demgegenüber unproblematisch sein. Denn durch die Möglichkeit, im zweiten Schritt die Entscheidung durch einen menschlichen Richter herbeizuführen, wird dem Anspruch auf rechtliches Gehör ausreichend Rechnung getragen.

Fraglich ist allerdings erneut, ob Art. 103 Abs. 1 GG den rein unterstützenden Einsatz künstlich intelligenter Systeme zulässt? Grundsätzlich dürften auch hier wenig Bedenken bestehen. Allerdings müsste auch unter dem Gesichtspunkt des rechtlichen Gehörs sichergestellt werden, dass Richter nicht dem *automation bias* unterliegen und die Ergebnisse der Maschine nicht automatisch und ohne Berücksichtigung der Umstände des Einzelfalls blind übernehmen.[86] Auch insofern greift deshalb das Postulat, Richter frühzeitig auf den Umgang mit KI-Systemen vorzubereiten und Verantwortlichkeiten zu verdeutlichen.

c) Richterliche Unabhängigkeit (Art. 97 GG)

Eine letzte hier zu thematisierende rechtliche Herausforderung künstlich intelligenter Systeme betrifft schließlich die richterliche Unabhängigkeit. Sie wird durch Art. 97 Abs. 1 GG garantiert und schützt den Richter – und damit auch die rechtssuchenden Parteien – vor einer inhaltlichen Einflussnahme von außen, namentlich durch den Staat und die Gesellschaft.[87] Für den Einsatz von KI-Systemen bedeutet

---

[83] Ebenso – ganz allgemein – ganz allgemein – die *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 171.

[84] BVerfG NJW 1991, 2823 f.; Maunz/Dürig/*Remmert*, Grundgesetz, 88. EGL, Stand August 2098, Art. 103 Rn. 62 ff.; BeckOK/*Radtke*/*Hagemeier*, GG, 41. Ed., 15.5.2019, Art. 101 Rn. 7 ff.

[85] BVerfG NJW 1991, 2823 f.

[86] *Enders*, JA 2018, 721 (723).

[87] Maunz/Dürig/*Hillgruber*, Grundgesetz, 88. EL, Stand August 2019, Art. 97 Rn. 21 ff.; BeckOK GG/*Morgenthaler*, 41. Ed., 15.2.2019, Art. 97 Rn. 3.

dies zunächst einmal, dass es keinen Zwang zur Nutzung geben darf.[88] Schon gar nicht dürfen Richter gezwungen werden, einem eventuellen Entscheidungsvorschlag zu folgen. Denn Richter sind in Deutschland nach Art. 97 Abs. 1 GG nur dem Gesetz unterworfen und – anders als Richter in den Ländern des *common law* – an Entscheidungen anderer – auch übergeordneter – Gerichte nicht gebunden.[89]

Fraglich ist allerdings, ob die bloße Verfügbarkeit künstlich intelligenter Systeme – und die Möglichkeit der freiwilligen Nutzung – den Grundsatz der richterlichen Unabhängigkeit beeinträchtigen könnte? Bestünde unter Umständen die Gefahr, dass sich Richter – *de facto* – unter Druck gesetzt sehen, die einschlägigen Systeme zu benutzen und Entscheidungsvorschlägen zu folgen, weil sie befürchten, ansonsten zeitlich oder inhaltlich ins Hintertreffen zu geraten? Auch diese Gefahr ist nicht von der Hand zu weisen.[90] Allerdings kann die Lösung auch hier nicht darin bestehen, Richtern eine potentielle Erkenntnisquelle – dieses Mal zu ihrem eigenen Schutz – vorzuenthalten.[91] Die richtige Reaktion kann vielmehr auch hier nur sein, Richter auf einen kritischen Umgang mit künstlich intelligenten System zu verpflichten und keinen Zweifel daran zu lassen, dass sie allein – und nicht das verwendete System – für Entscheidungen verantwortlich sind.[92]

Eine Grenze des Einsatzes künstlich intelligenter System dürfte unter dem Gesichtspunkt der richterlichen Unabhängigkeit freilich dort zu ziehen sein, wo Richter eine unbotmäßige Kontrolle ihrer Arbeit und deshalb eine Beeinträchtigung ihrer Entscheidungsfreiheit befürchten.[93] Frankreich hat deshalb im Jahr 2019 ein Gesetz verabschiedet, dass es verbietet, Informationen über einzelne Richter zu sammeln und systematisch auszuwerten.[94] Ob dies der richtige Weg ist, soll an dieser Stelle dahingestellt bleiben. Wichtig ist aber, dass der Einsatz von KI-Systemen nicht dazu führen darf, dass die sachliche Unabhängigkeit von Richtern eingeschränkt wird.

---

[88] Ebenso *Enders*, JA 2018, 721 (723).

[89] BVerfG NJW 1988, 2787; BVerfG NJW 1993, 996; Maunz/Dürig/*Hillgruber*, Grundgesetz, 88. EL, Stand August 2019, Art. 98 I Rn. 49; BeckOK GG/*Morgenthaler*, 41. Ed., 15.2.2019, Art. 97 Rn. 11.

[90] Ebenso die Einschätzung der *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), 46 sowie der Datenethikkommission der Bundesregierung, Gutachten Datenethikkommssion, 2019, 213.

[91] Anders wohl die Einschätzung der *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommssion, 2019, 213, 218.

[92] Ebenso – ganz allgemein – die Empfehlungen der *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 166 f.

[93] So auch die Einschätzung von *Fries*, NJW 2016, 2860 (2864), der *European Commission for the Efficiency of Justice*, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, CEPEJ(2018), S. 11 sowie der *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 213. A. A. wohl *Römermann*, LTO v. 2. Januar 2020.

[94] Article 33 II 1 Loi no. 2019-222 du 23 mars 2019 de programmation 2018–2022 et de réforme pour la justice. Siehe dazu France's Controversial Judge Data Ban – The Reaction, Artificial Lawyer v. 5. Juni 2019, abrufbar unter https://www.artificiallawyer.com/2019/06/05/frances-controversial-judge-data-ban-the-reaction (zuletzt abgerufen am 10.12.2021); *Kuhlmann*, LTO v. 14. Juni 2019.

IV. Fazit und Ausblick

Die vorstehenden Überlegungen zeigen, dass der Einsatz von KI-Systemen in der Justiz große Chancen bietet, aber auch zahlreiche Fragen aufwirft. Diese müssen zufriedenstellenden beantwortet werden, bevor Künstliche Intelligenz tatsächlich für die zivilgerichtliche Streitbeilegung nutzbar gemacht werden kann. Insbesondere müssen die für den Einsatz in der Justiz erforderlichen Anforderungen definiert und bei der (Weiter-) Entwicklung der entsprechenden Systeme berücksichtigt werden.[95] Zu begrüßen ist deshalb, dass der Vorschlag für eine KI-Verordnung, den die Europäische Kommission im April 2021 vorgelegt hat,[96] KI-Systeme, die bei der Ermittlung und Auslegung von Sachverhalten und Rechtsvorschriften zum Einsatz kommen sowie bei der Anwendung des Rechts auf konkrete Sachverhalte unterstützen sollen, den Hochrisiko-KI-Systemen i. S. v. Titel III zuordnet,[97] die besonders strengen Anforderungen unterliegen.[98] Sollte der Entwurf verabschiedet werden, wird der Einsatz von KI-Systemen in der Justiz deshalb nur dann zulässig sein, 1) wenn sie mit Datensätzen entwickelt wurden, die relevant, repräsentativ, fehlerfrei und vollständig sind,[99] 2) wenn sie so konzipiert sind, dass sie im Hinblick auf ihre Zweckbestimmung ein angemessenes Maß an Genauigkeit und Robustheit aufweisen,[100] 3) wenn sie über eine Protokollierungsfunktion verfügen, die gewährleistet, dass das Funktionieren des Systems in einem der Zweckbestimmung des Systems angemessenen Maße rückverfolgbar ist,[101] und 4) wenn ihr Betrieb hinreichend transparent ist, damit die Nutzer die Ergebnisse des Systems angemessen interpretieren und verwenden können[102] Da Hochrisiko-KI-System zudem erst nach Durchlaufen eines Konformitätsverfahrens in Verkehr gebracht[103] und nur unter menschlicher Aufsicht eingesetzt werden dürfen,[104] trägt der Vorschlag der Europäischen Kommission zumindest

---

[95] Die Diskussion über die Einsatzbedingungen von künstlicher Intelligenz in der Justiz ist insofern Teil der allgemeinen Diskussion über die Regulierung künstlicher Intelligenz. Siehe dazu nur *Busch*, Algorithmic Accountability, 2018; *Datenethikkommission der Bundesregierung*, Gutachten Datenethikkommission, 2019, 159 ff.; *Hochrangige Expertengruppe für künstliche Intelligenz*, Ethikleitlinie für eine vertrauenswürdige KI 2018.

[96] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union, COM(2021) 206 final. Siehe dazu die ersten Einschätzungen von *Geminn*, ZD 2021, 354; *Heiss*, NZG 2021, 611; *Lachenmann/Meyer*, MMR-Aktuell 2021, 438173; *Nelles*, ZD-Aktuell 2021, 05194; *Valta/Vasel*, ZRP 2021, 142.

[97] Art. 6 Abs. 2 KI-VO-Entwurf i. V. m. Anhang III Nr. 8a).

[98] Art. 8 ff. KI-VO-Entwurf.

[99] Näher dazu Art. 10 KI-VO-Entwurf.

[100] Näher dazu Art. 15 KI-VO-Entwurf.

[101] Näher dazu Art. 12 KI-VO-Entwurf.

[102] Näher dazu Art. 13 KI-VO-Entwurf.

[103] Näher dazu Art. 19 KI-VO-Entwurf.

[104] Näher dazu Art. 14 KI-VO-Entwurf.

einigen der oben dargestellten Probleme beim Einsatz datengestützter Entscheidungsvorhersagesysteme in der Justiz Rechnung. Sollte der Entwurf für eine KI-Verordnung Gesetz werden, wird damit hoffentlich die Grundlage dafür gelegt werden, dass sich in einigen Jahren auch deutsche Richter von Entscheidungsvorhersagesystemen unterstützen lassen können.

Vertrauenswürdige Künstliche Intelligenz

Nachvollziehbar, Transparent, Korrigierbar

Ute Schmid

I. Einleitung

Mit der digitalen Transformation halten auch Anwendungen der Künstlichen Intelligenz (KI) in immer mehr Arbeits- und Lebensbereiche Einzug. Insbesondere kommen immer mehr aus Daten gelernte Modelle zum Einsatz, die meistens intransparente Black-boxes sind. Dass Menschen nachvollziehen können, warum ein KI-System sich wie verhält, ist aus verschiedenen Gründen notwendig: Die Modellentwicklerinnen und -entwickler selbst müssen in der Lage sein, Eigenschaften der gelernten Modelle zu beurteilen – insbesondere auch mögliche Biases aufgrund von Überanpassung an die zum Lernen genutzten Daten. Für sicherheitskritische Anwendungen werden zunehmend auch Aspekte der Zertifizierung und Prüfung relevant. Domänenexpertinnen und -experten – etwa in der medizinischen Diagnostik oder bei der Qualitätskontrolle in der industriellen Produktion – müssen Systementscheidungen nachvollziehen, überprüfen, und gegebenenfalls auch korrigieren können. Verbraucherinnen und Verbraucher sollten verstehen, warum sich ein System – eine Smart Home Steuerung, eine Fahrassistenz – auf eine bestimmte Art verhält und warum ihnen bestimmte Produkte empfohlen, bestimmte Tarife angeboten oder bestimmte Angebote vorenthalten werden. Im Beitrag wird nach einer kurzen Einführung in das Thema KI ein Überblick über Methoden der sogenannten dritten Welle der KI gegeben. Zentral sind hier Ansätze der sogenannten erklärbaren KI (*eXplainable AI*, XAI), die ermöglichen sollen, dass die Entscheidungen von KI-Systemen nachvollziehbar werden. Die wesentlichen Ansätze werden charakterisiert und aufgezeigt für welche Zielsetzungen und Anwendungen sie jeweils geeignet sind. Es wird aufgezeigt, dass neben den vielbeachteten Methoden zur Visualisierung besonders auch Methoden wichtig sind, die erlauben, Systementscheidungen differenziert zu beschreiben. Zudem wird argumentiert, neben der Nachvollziehbarkeit auch Interaktivität und Korrigierbarkeit von KI-Systemen notwendig sind, damit KI-Systeme menschliche Kompetenzen nicht einschränken sondern partnerschaftlich unterstützen.

II. Wissensbasierte und datengetriebene Ansätze der Künstlichen Intelligenz

Künstliche Intelligenz ist ein Forschungsgebiet der Informatik, in dem Algorithmen zur Lösung von Problemen entwickelt werden, die Menschen im Moment noch besser lösen.[1] Das Gebiet erhielt seine Bezeichnung ‚Artificial Intelligence' im Jahr 1956 vom Informatik-Pionier John McCarthy an der Universität Stanford. Die beiden wichtigsten Familien von KI-Methoden sind wissensbasierte Methoden und maschinelles Lernen.[2] Beide Gebiete wurden von Beginn an betrachtet. Die erste Umsetzung eines Programms zum maschinellen Lernen war ein Programm zum Lernen einer Strategie für das Dame-Spiel und wurde von Arthur Samuel im Jahr 1952 realisiert. Zu den frühen Ansätzen gehörten auch das Perzeptron als Modell eines einzelnen Neurons sowie Entscheidungsbaum-Algorithmen.

In den 1980er Jahren war die Hochphase für wissensbasierte Methoden im Kontext von Anwendungen für Expertensysteme, Man erhoffte sich, dass KI-Systeme menschliche Experten in vielen Bereichen entlasten oder unterstützen könnten – von der medizinischen Diagnostik über die Planung von Produktionsprozessen bis zur Nutzung intelligenter Tutorsysteme im Unterricht. Im Kontext der Forschung zu wissensbasierten Systemen entstanden effiziente Algorithmen zum Ziehen von Schlussfolgerungen. Es wurden spezielle KI-Programmiersprachen sowie spezifische Hardware für effizientere Verarbeitung, insbesondere die Lisp Machine. Forschung zu maschinellem Lernen fand zwar nach wie vor statt, wurde aber von den wissensbasierten Ansätzen dominiert. Die Hochphase der Expertensysteme weist entsprechend Gemeinsamkeiten zum gegenwärtigen Hype im Bereich maschinelles Lernen auf. Wieder dominiert eine Richtung stark und für tiefe neuronale Netze werden spezielle Programmbibliotheken sowie spezielle Hardware in Form von GPUs (*Graphics Processing Units*), die besonders effizient Matrizen multiplizieren können, entwickelt.

Die großen Hoffnungen, die in Expertensysteme gesetzt wurden, konnten letztendlich nur teilweise erfüllt werden, insbesondere aufgrund des sogenannten *Knowledge Engineering Bottleneck* – der Erkenntnis, dass menschliches Wissen nur in Teilen explizit verfügbar ist und formal repräsentiert werden kann. Große Bereiche menschlichen Wissens, vor allem perzeptuelles Wissen und hochautomatisierte Handlungsroutinen sind implizit und können nicht oder nur unzureichend mit Methoden der Wissensakquisition erfasst werden. Das Phänomen wird auch als Polanyis Paradox bezeichnet: Wie können wir Menschen mehr wissen, als das, worüber wir reden können?

Beeindruckend Erfolge in der Anwendung von tiefen neuronalen Netzen haben seit etwa 2010 eine neue Hochphase der KI eingeläutet – diesmal mit Fokus auf

---

[1] *Rich*, Artificial Intelligence, 1983, McGraw-Hill.

[2] Siehe dazu: *Russel/Norvig*, Artificial Intelligence A Modern Approach, 4. Aufl. 2020, Pearson.

maschinellem Lernen. Wesentlich an dem neuen großen Interesse an KI ist, dass es erstmals möglich war, aus verschiedenen Arten von Daten, wie Bildern oder Texten nahezu direkt, ohne aufwendige Vorverarbeitung zu lernen (*end-to-end learning*). Die meisten Ansätze des maschinellen Lernens, auch die klassischen neuronalen Netze, wie sie seit Ende der 1980er Jahre entwickelt wurden, erwarten als Eingabe Daten in Form von Merkmalsvektoren. Viele Daten liegen ohnehin in Tabellenform vor – beispielsweise Kundendaten oder Patientendaten. Möchte man aber beispielsweise aus Bilddaten wie Fotos von Objekten oder auch Röntgenbildern lernen, muss man für die klassischen Ansätze des maschinellen Lernens zunächst Merkmale wie Texturen oder Farbverteilungen aus den vorliegenden Bilddaten extrahieren. Genau wie für die wissensbasierten Ansätze der KI stellten perzeptuelle Aufgaben auch für maschinelles Lernen eine Herausforderung dar.

Im Jahr 2012 gewann erstmalig ein tiefes neuronales Netz – ein *Convolutional Neural Network* (CNN) namens AlexNet – bei der ImageNet Challenge.[3] Bei der Challenge sollen Bilder aus 1000 Kategorien, beispielsweise Tierarten, Fahrzeugtypen, Gebäude, klassifiziert werden. Dazu stehen mehrere Millionen an Bildern zur Verfügung, für die mit Hand annotiert wurde, welche Objekte darauf abgebildet sind. Anders als frühere Ansätze des maschinellen Lernens konnte das AlexNet direkt aus den Bildern lernen. Vergleichbare Entwicklungen gibt es für die Verarbeitung natürlicher Sprache, etwa maschinelle Übersetzung (DeepL) oder Textgenerierung (GPT3). Wieder sind allerdings die Erwartungen an das, was diese neuartigen KI-Methoden leisten können, überzogen. Polanyis Rache[4] hat das Pendel von einem nahezu exklusiven Fokus auf KI-Methoden für explizites Wissen auf einen alleinigen Fokus auf KI-Methoden für implizites Wissen schwingen lassen. Für jedes Problem wird Lernen aus vielen Daten als der einzig sinnvolle Zugang angesehen. Vorhandenes Wissen, inklusive sorgfältig gewonnenes Wissen über Kausalbeziehungen, wird über Bord geworfen, um Dinge unperfekt aus Daten zu lernen für die explizites Wissen verfügbar ist. Gleichzeitig verzichtet man auf Nachvollziehbarkeit und Kontrolle, da tiefe neuronale Netze Eingaben auf komplexe Art mathematisch verrechnen und damit Blackboxes sind.

III. Probleme mit datenintensivem Maschinellen Lernen

Auch wenn datenintensives maschinelles Lernen mit der neuen Generation von tiefen neuronalen Netzen für verschiedene Anwendungsbereiche neue Möglichkeiten eröffnet, bringt es auch neue Probleme mit sich. Die Voraussetzungen an

---

[3] *Krizhevsky/Sutskever/Hinton*, Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, NeurIPS 2012, S. 1097–1105.

[4] *Kambhampati*, Polanyi's Revenge and AI's New Romance with Tacit Knowledge, Communication of the ACM, 64 (2), 2021, S. 31–32.

Menge und Qualität von Daten ist extrem hoch. Das bereits genannte ImageNet besteht aus 14 Millionen Bildern und 20.000 Kategorien. Oft wird übersehen, dass der Aufwand, Wissen zu erfassen und für die Verarbeitung durch KI-Methoden formalisieren, beim maschinellen Lernen nicht verschwindet, sonder auf das korrekte Annotieren von Trainingsdaten verschoben wird. Click-worker müssen jedes Beispiel von Hand mit der korrekten Kategorie versehen – oder auch Objekte in Bildern markieren. Je komplexer die Architektur eines neuronalen Netzes ist, desto mehr Daten werden benötigt, um es zu trainieren. Sind zu wenige Daten vorhanden, so werden diese vervielfältigt (augmentiert). Bilder werden beispielsweise in ihren Farbwerten verändert. In komplexen Anwendungsbereichen, bei denen unklar ist, welche komplexe Kombination von Informationen verantwortlich für eine bestimmte Kategorie ist, kann dies zu unerwünschten Verzerrungen (*biases*) führen. Beispielsweise wird bei der Diagnose von Tumoren aus Gewebeschnitten das Gewebe häufig eingefärbt. Ein Modell, das entscheidet, ob und wenn ja welche Kategorie eines Tumors vorliegt, könnte durch Trainingsdaten mit anderen Färbungen als den originalen in die Irre geführt werden.

Überwachte Ansätze des maschinellen Lernens, und dazu gehören auch viele Ansätze von tiefen neuronalen Netzen, benötigen eine für das Problem möglichst repräsentative Stichprobe von Trainingsdaten, die mit der korrekten Ausgabe annotiert sind – man spricht von *ground truth labeling*. Gerade in der Medizin, aber auch in anderen Anwendungsbereichen, ist häufig nicht eindeutig klar, was die korrekte Entscheidung für ein gegebenes Datum ist. So könnte es sein, dass ein Medizinexperte oder eine Medizinexpertin für dieselbe Aufnahme eines Gewebeschnitts auf Tumorklasse pT3 entscheidet, ein oder eine andere auf pT4. Fehlen bestimmte Arten von Daten in der Trainingsmenge (*sampling bias*) und sind Daten nicht korrekt annotiert hat das direkte Auswirkung auf die Qualität des gelernten Modells.[5] Zusätzlich gilt, dass Modelle, die aus Daten generiert werden, typischerweise nur für ähnliche Daten, die innerhalb der Verteilung der Daten der Trainingsmenge liegen, generalisieren können, aber nicht für Daten, die außerhalb der Verteilung liegen. Hat man ein Modell trainiert, das Autoarten unterscheiden kann und es erhält später eine Waschmaschine als Eingabe, so wird es diese bezüglich der Ähnlichkeit zu den gelernten Autoarten klassifizieren. Ein Mensch würde dagegen sagen, das ist ja etwas ganz anderes, als ich bisher gesehen habe, dazu kann ich nichts sagen. Gelernte Modelle verfügen standardmäßig nicht über eine solche Art der Meta-Kognition. Bei einem wissensbasierten KI-System würde dagegen eine Eingabe außerhalb des betrachteten Bereichs nicht bearbeitet. Die Güte gelernter Modelle hängt also stark von der Auswahl und Qualität der Daten ab, mit denen es trainiert wurde.

---

[5] Siehe *Bruckert/Finzel/Schmid*, The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions, Frontiers in Artificial Intelligence 3, 2020, 507973.

Aber auch wenn die Daten repräsentativ gesammelt und korrekt annotiert sind, kann es zu unerwünschten Effekten kommen. Unfairheit in der Realität wird in den Daten repräsentiert. Wenn in einer Firma deutlich weniger Frauen im Bereich IT arbeiten als Männer und man trainiert ein Modell zur Bewerbungsauswahl naiv einfach mit den vorhandenen Daten, so führt das dazu, dass eine Bewerberin überhaupt nicht mehr für eine Position in der IT berücksichtigt wird, wie beim Recruiting Tool von Amazon 2018 geschehen.[6] Sind einem solche unfairen Verteilungen in den Daten im Vorhinein bewusst, so kann dies durch entsprechende Methoden im Lernprozess berücksichtigt werden. Generell lassen sich unfaire Modelle aber nicht ausschließen.

Menschliches wie maschinelles Lernen ist ein Schließen von einer Stichprobe von Daten oder Erfahrungen auf eine Grundgesamtheit. Solche induktiven Schlüsse können nie vollständig korrekt sein. Menschlicher Konzepterwerb ist im Allgemeinen sehr robust. So gelingt es uns ohne Probleme, Katzen von anderen Tieren zu unterscheiden, auch bei ganz unterschiedlichen Arten von Katzen, Beleuchtungen oder Hintergründen. In anderen Bereichen neigen Menschen zur Übergeneralisierung und bilden Stereotype und Vorurteile aus. Vorurteile, die sich auf Geschlecht oder Ethnie beziehen lassen sich zwar nicht ausschließen, aber sie können erkannt und auch korrigiert werden. Aber bei menschlichem wie maschinellem Lernen gilt, dass man Fehler machen kann. Bei maschinell gelernten Modellen schätzt man ab, welche Fehlerrate es für ungesehene Daten aufweisen wird. Eine Vorhersagegenauigkeit (*predictive accuracy*) von 99 % klingt nicht schlecht, heißt aber, dass das Modell bei jeder hundertsten Eingabe einen Fehler macht. Sucht man mit einer Suchmaschine nach Bildern von Katzen, so ist es nicht schlimm, wenn auf jedem hundertsten Bild etwas anderes zu sehen ist. Die Vorteile überwiegen. Man schaut sich die Bilder an und sucht sich ein geeignetes aus. Würde bei einer medizinischen Diagnose bei jedem hundertsten Fall irrtümlich eine Krankheit diagnostiziert oder – noch schlimmer – übersehen, wäre das untragbar. Desgleichen ist es sicher nicht wünschenswert, dass jeder hundertsten Person ein Kredit fälschlicherweise verwehrt oder ein Versicherungstarif grundlos zu hoch angesetzt wird.

Um solche unerwünschten Modellentscheidungen zu erkennen und korrigieren zu können, muss nachvollziehbar sein, aufgrund welcher Information das Modell zu seiner Entscheidung kam. Allerdings liefern viele Ansätze des maschinellen Lernens, insbesondere auch tiefe neuronale Netze, intransparente Modelle, die auch für die Modellentwickelnden selbst Black-boxes sind.

---

[6] *Dastin*, Amazon scraps secret AI recruiting tool that showed bias against women, 11. Oktober 2018, Reuters, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (abgerufen am 28.2.2022)

IV. Erklärbare Künstliche Intelligenz – Nachvollziehbarkeit von maschinell gelernten Modellen

Nachdem seit etwa 2015 in immer mehr Anwendungsbereichen das Interesse am Einsatz von datenintensiven KI-Methoden wuchs, wurde schnell deutlich, dass ein ausschließlicher Fokus auf Black-box Ansätze des maschinellen Lernens oft weder möglich noch wünschenswert ist. Anwendungsmöglichkeiten werden durch die oben diskutierten Anforderungen an die Datenmenge und Qualität, insbesondere aber durch die hohen Aufwand an die Annotation der Trainingsdaten, eingeschränkt. Zudem wurde schnell deutlich, dass vor allem in sicherheitskritischen Bereichen wie etwa der Medizin Systeme, bei denen nicht nachvollziehbar ist, auf welcher Grundlage sie zu einer Entscheidung oder einer Handlungsempfehlung kommen, nicht akzeptabel ist. In Bereichen, die direkten Einfluss auf Verbraucherinnen und Verbraucher haben – von personalisierter Werbung bis zu Kreditvergabe – wurde ebenfalls bald das Recht aus Transparenz eingefordert (Goodman & Flaxman, 2017).[7]

Im Frühjahr 2017 wurde von der DARPA (*Defense Advanced Research Projects Agency, U.S.A.*) das Explainable Artificial Intelligence Programm gestartet. Ziel des Programms ist die Entwicklung von Methoden, die (a) zu maschinell gelernten Modellen führen, die besser nachvollziehbar sind als Black-box-Modelle aber gleichzeitig ein hohes Maß an Vorhersagegenauigkeit behalten, und (b) Anwenderinnen und Anwendern ermöglichen, diese neu entstehende Generation von partnerschaftlichen KI-Systemen zu verstehen, den Entscheidungen angemessen zu vertrauen und effektiv mit den Systemen zu interagieren.[8] Beispielhaft wurde an der Klassifikation einer Katze durch ein neuronales Netz gezeigt, dass eine Erklärung der Modellentscheidung sowohl verbalisierbare Merkmale wie „hat Fell, Schnurrhaare und Krallen" als auch prototypische Bilder typischer visueller Merkmale wie die Form der Ohren beinhalten kann.[9] Der Begriff ‚*explainable*'/‚erklärbar' führte jedoch außerhalb der Forschung zu Missverständnissen, da eher suggeriert wird, dass die Arbeitsweise von KI-Systemen für Laien verständlich erklärt wird und nicht, dass ein KI-System, speziell ein maschinell gelerntes Modell, mit einer Schnittstelle verstehen wird, die Systementscheidungen nachvollziehbar macht. Parallel wurden zu Beginn auch Bezeichnungen wie ‚*comprehensible machine learning*'[10] (nachvollziehbares maschinelles Lernen) oder

---

[7] *Goodman/Flaxman*, European Union regulations on algorithmic decision-making and a „right to explanation", AI Magazine, 38 (3), 2017, S. 50–57.

[8] *Gunning/Aha*, DARPA's explainable artificial intelligence (XAI) program, AI magazine, 40 (2), 2019, S. 44–58.

[9] Siehe https://twitter.com/darpa/status/843067035366187008, 18.3.2017.

[10] *Schmid*, Inductive Programming as Approach to Comprehensible Machine Learning, In Proceedings of DKB/KIK@KI, 2018, S. 4–12.

interpretierbares maschinelles Lernen[11] genutzt. Inzwischen wird häufig auch von erklärendem (‚*explanatory*') maschinellem Lernen gesprochen.[12] Transparenz wird in Zusammenhang mit KI-Systemen meist allgemeiner aufgefasst als Erklärbarkeit: Es soll deutlich gemacht werden, wenn einer Empfehlung oder Entscheidung auf der Nutzung von KI-Methoden basiert oder auch, dass eine Interaktion nicht mit einem Menschen sondern einem KI-System wie einem Chatbot erfolgt.

Inzwischen hat sich eine Vereinheitlichung der Begrifflichkeiten entwickelt: Nach dem anfänglichen Fokus auf Erklärbarkeit für tiefe neuronale Netze, wird inzwischen die Relevanz von Methoden zur Generierung von Erklärungen, kurz XAI-Methoden, für alle Arten von KI-Systemen gesehen. Zum einen werden Erklärungsmethoden für verschiedene Black-box Ansätze des maschinellen Lernens entwickelt (hierzu gehören auch Methoden wie *Support Vector Machines* oder k-nächste Nachbarn Ansätze)[13]. Zum anderen werden Erklärungsmethoden auch für wissensbasierte KI-Systeme sowie für *White-box* Ansätze des maschinellen Lernens entwickelt. Für diese Systeme ist zwar im Prinzip nachvollziehbar, wie eine Entscheidung zustande kommt. Aber – vergleichbar mit großen Software-Systemen – sind die Modelle oft zu komplex, um den gesamten Prozess der Informationsverarbeitung zu durchblicken. Außerdem sind die Modelle in speziellen Repräsentationsformalismen gespeichert, die für die Verarbeitung durch Computerprogramme ermöglichen und müssen entsprechend geeignet in nachvollziehbare Erklärungen übersetzt werden.

Inzwischen hat sich eingebürgert, dass White-box Ansätze des maschinellen Lernens, wie beispielsweise Entscheidungsbaumverfahren als interpretierbares maschinelles Lernen bezeichnet werden.[14]

Inzwischen existiert eine große Bandbreite von XAI-Methoden, die sich für verschiedene Zielgruppen und verschiedene Informationsbedarfe eignen. Es gibt zahlreiche Methoden, die die Relevanz von spezifischen Informationen aus der Eingabe für die aktuelle Entscheidung aufzeigen. Dies können Merkmale, Worte oder Teile von Bildern sein. Beispielsweise zeigt der Ansatz LIME,[15] welche Gruppen von Bildpunkten für eine Klassifikationsentscheidung vorhanden sein müs-

---

[11] *Doshi-Velez/Kim*, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint, arXiv: 1702.08608.

[12] *Ai/Muggleton/Hocquette/Gromowski/Schmid*, Beneficial and harmful explanatory machine learning, Machine Learning, 110 (4), 2021, S. 695–721; *Teso/Kersting*, Explanatory interactive machine learning, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, S. 239–245.

[13] Siehe für eine allgemeinverständliche Einführung z. B.: *Kersting/Lampert/Rothkopf*, Wie Maschinen lernen: Künstliche Intelligenz verständlich erklärt, 2019, Springer.

[14] *Rudin*, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 2019, S. 206–215.

[15] *Ribeiro/Singh/Guestrin*, „Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, S. 1135–1144.

sen – etwa, dass Auge und Ohr relevant dafür ist, ob das Modell eine Katze erkennt. LIME ist ein sogenannter modell-agnostischer Erklärungsansatz: Zur Erklärungsgenerierung wird nicht in das gelernte Modell eingegriffen, stattdessen werden die Eingabedaten manipuliert und die sich daraus ergebende Modellentscheidung betrachtet. Ein Ansatz, der speziell für Bildklassifikation mit (tiefen) neuronalen Netzen entwickelt wurde, ist LRP (*Layerwise Relevance Propagation*).[16] Hier werden diejenigen Bildpunkte hervorgehoben, die besonders starken Einfluss auf die Ausgabe des Netzes hatten. Im Gegensatz zu LIME ist LRP modell-spezifisch, das heißt, die Methode muss direkt in den Lernalgorithmus integriert werden. Das Hervorheben der Information, die für ein gelerntes Modell besonders relevant ist, ist besonders für die Modellentwicklerinnen und -entwickelt nützlich, um zu prüfen, ob das Modell sinnvoll generalisiert hat. Beim Lernen kann es nämlich passieren, dass das Modell irrelevante Information zur Vorhersage nutzt, die mit der vorherzusagen Klasse korreliert. Das Modell passt sich also zu stark an die Trainingsdaten an (*overfitting*), was zu Problemen bei der Vorhersage für noch nie gesehen Daten führen kann. Man spricht hier auch von „*right for the wrong reasons*" oder von „Kluge Hans"-Prädiktoren. Beispielsweise könnte es sein, dass zufällig ein Teil der Fotos, auf denen Pferde zu sehen sind, mit eine Quellenangabe (z. B. eine Webseite) angegeben ist. Der Lernalgorithmus kann dann diese einfachere Information nutzen, um für die vorhandenen Daten korrekt anzugeben, wann ein Pferd zu sehen ist. Dass diese Ausgabe aber auf Basis der Quellenangabe erfolgt, kann das Hervorheben der genutzten Bildpunkte zeigen.[17]

Für Domänenexpertinnen und -experten und auch für Endnutzende ist das ausschließliche Hervorheben von relevanten Informationen meist wenig hilfreich. So kann visuelles Hervorheben zwar zeigen, dass auf einem Gewebeschnitt tatsächlich ein bestimmter Tumor zu sehen ist. Um aber nachvollziehen zu können, warum das Modell auf Tumorklasse pT3 und nicht auf pT4 entschieden hat, sind deutlich komplexere Informationen notwendig, die besser sprachlich ausgedrückt werden können. Dazu gehören räumliche Relationen, wie die Lage des Tumors relativ zu anderen Gewebearten oder die konkrete Ausprägung einzelner Merkmale, etwa der Durchmesser des Tumors.[18] Solche Erklärungen können beispielsweise generiert werden, Black-Box Ansätze des maschinellen Lernens und interpretierbare Ansätze kombiniert werden.[19]

---

[16] *Bach*/*Binder*/*Montavon*/*Klauschen*/*Müller*/*Samek*, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one, 10 (7), 2015, e0130140.

[17] *Lapuschkin*/*Wäldchen*/*Binder*/*Montavon*/*Samek*/*Müller*, Unmasking Clever Hans predictors and assessing what machines really learn, Nature communications, 10 (1), 2019, S. 1–8.

[18] *Bruckert*/*Finzel*/*Schmid* (Fn. 5); *Schmid*, Interactive Learning with Mutual Explanations in Relational Domains, in: Muggleton/Chater (Hrsg.), Human-Like Machine Intelligence (chap. 17), 2021, S. 338–354, Oxford University Press.

[19] *Rabold*/*Deininger*/*Siebers*/*Schmid*, Enriching visual with verbal explanations for relational concepts -combining LIME with Aleph, in: Workshops at Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2019, S. 180–192, Springer.

Für Verbraucherinnen und Verbraucher sind häufig einfache Erklärungen wie man sie von Empfehlungssystemen kennt, relevant.[20] Wird einem etwa in einem Online-Shop ein bestimmtest Produkt empfohlen, so kann man nachfragen, auf welcher Datengrundlage diese Empfehlung getroffen wurde. Man erhält dann typischerweise frühere Käufe gezeigt, die für einen Ähnlichkeitsvergleich mit den Kaufprofilen anderer Personen abgeglichen wurden. Wenn es darum geht, dass transparent gemacht wird, wie Algorithmen (mit und ohne KI-Komponenten) bei Banken, Versicherungen oder andere Unternehmen zu bestimmten Entscheidungen, etwa der Ablehnung eines Kredits oder der Höhe eines Versicherungsbeitrags kommen, sind besonders kontrafaktische Erklärungen hilfreich[21] – beispielsweise:

„Sie haben den Kredit nicht erhalten, weil Ihr Jahreseinkommen 45.000 € beträgt. Wenn Ihr Jahreseinkommen 55.000 € betragen würde, hätten Sie den Kredit bekommen."

Solche Erklärungen geben die relevante Information an Kunden weiter, wobei gleichzeitig vermieden wird, dass Unternehmen ihre Algorithmen preisgeben müssen.

Eine weitere Möglichkeit für Erklärungen liefern prototypische sowie kontrastive Beispiele. Solche Beispiele bieten insbesondere für Expertinnen und Experten die Möglichkeit, besser zu verstehen, wie das Modell strukturiert ist. Die bisher betrachteten XAI-Methoden erklären, wie eine konkrete Entscheidung zustande kam (lokale Erklärung). Speziell ausgewählte Beispiele können aufzeigen, welche Daten ein Modell als besonders typisch für eine bestimmte Klasse beurteilt und welche Daten knapp an den Entscheidungsgrenzen liegen. Erklärungen, die das gesamte Modell betreffen, werden auch als globale Erklärungen bezeichnet. Kontrastive Beispiele können als *near miss* Erklärungen auch als lokale Erklärungen genutzt.[22] Für einen aktuellen Fall – beispielsweise einen Gewebeschnitt, der als Tumorklasse pT3 klassifiziert wird – kann der ähnlichste Fall gezeigt werden, der aber in eine andere Klasse, etwa pT4, klassifiziert wird.

Erklärbare KI besteht also aus einer wachsenden Menge verschiedener Methoden, die jeweils für verschiedene Informationsziele passen. Theoretische und empirische Analysen der Eigenschaften und Wirkung von Erklärungen aus der Psychologie fließen zunehmend in die Forschung zu XAI ein.[23] XAI-Methoden sind ein wichtiger Beitrag für die Nachvollziehbarkeit von KI-Systemen, insbesondere von maschinellem Lernen. Allerdings muss im jeweiligen Anwendungskon-

---

[20] *Tintarev/Masthoff*, Evaluating the effectiveness of explanations for recommender systems, User Modeling and User-Adapted Interaction, 22 (4), 2012, S. 399–439.

[21] *Wachter/Mittelstadt/Russell*, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology, 31 (2), 2017, S. 841.

[22] *Rabold/Siebers/Schmid*, Generating Contrastive Explanations for Inductive Logic Programming Based on a Near Miss Approach, Machine Learning, online first, Sept. 2021, https://doi.org/10.1007/s10994-021-06048-w■.

[23] *Miller*, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, 267, 2019, S. 1–38.

text, insbesondere im beruflichen Umfeld, sorgfältig geprüft werden, dass Erklärungen tatsächlich genutzt werden, um Systementscheidungen zu kontrollieren und sich damit nicht blindes, sondern gerechtfertigtes Vertrauen in ein KI-System entwickeln kann.[24] Die Gefahr besteht, dass das bloße Vorhandensein der Möglichkeit einer Erklärung dazu führt, dass die Systementscheidungen unreflektiert übernommen werden.[25]

V. KI-Methoden der dritten Welle: Hybrid, nachvollziehbar und korrigierbar

Methoden der Erklärbaren KI werden auch als Auftakt zur dritten Welle der KI bezeichnet – nach der ersten Welle der wissensbasierten Ansätze („*describe*"), folgte datenintensives maschinelles Lernen („*categorize*"), das abgelöst werden soll durch Ansätze, die sich kontextabhängig an die Interessen der Nutzenden anpassen („*explain*"). Zunehmend wird argumentiert, dass die für die dritte Welle notwendigen Methoden nicht nur das Generieren von Erklärungen adressieren müssen, sondern, dass maschinelles Lernen Interaktion, insbesondere Korrekturen des Modelles, erlauben sollte.[26] Zudem wird gesehen, dass eine Kombination aus wissensbasierten Ansätzen und maschinellem Lernen zu datensparsameren und robusteren Modellen führen kann.

Die Kombination von erklärendem und interaktivem maschinellen Lernen ist ein sinnvoller Ansatz, um den oben diskutierten Problemen mit der Menge und Qualität von Daten entgegenzuwirken. So können Expertinnen und Experten, eine Systementscheidung, die sie direkt nachvollziehen können, einfach akzeptieren, eine Systementscheidung genauer hinterfragen, indem sie eine oder auch mehrere Erklärungen vom System anfordern, wie die Entscheidung zustande kam, und im dritten Schritt, diese Entscheidung auch korrigieren. Während die meisten Arbeiten zu interaktivem maschinellen Lernen nur die Korrektur der Ausgabe ermöglichen, gibt es nun erste Ansätze, die zusätzlich die Korrektur der Erklärungen erlauben. Dadurch kann die Anpassung des Modells gezielt gesteuert werden.[27] Interaktion erlaubt also, dass gezielt menschliches Wissen in den Lernprozess eingebracht werden kann. Korrekturen sind auch dann möglich, wenn

---

[24] *Thaler*/*Schmid*, Explaining machine learned relational concepts in visual domains-effects of perceived accuracy on joint performance and trust, in: Proceedings of the Annual Meeting of the Cognitive Science Society, 43 (43), 2021, S. 1705–1711.

[25] *Lee*/*See*, Trust in automation: Designing for appropriate reliance, Human Factors, 46 (1), 2004, S. 50–80.

[26] *Teso*/*Kersting* (Fn. 12); *Müller*/*März*/*Scheele*/*Schmid*, An Interactive Explanatory AI System for Industrial Quality Control, Thirty-Fourth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-22, Feb 22 – March 1 2022, virtual colocated with 36 AAAI Conference on Artificial Intelligence), 2022.

[27] *Schmid* (Fn. 18).

Wissen nicht vollständig explizit gemacht werden kann. So kann beispielsweise ein Experte oder eine Expertin durchaus erkennen, ob eine Diagnose akzeptabel ist oder nicht und eventuell auch bei deren Begründung fehlerhafte Annahmen identifizieren. Gleichzeitig kann davon ausgegangen werden, dass die Möglichkeit zur Korrektur, zu einem stärkeren Gefühl von Kontrolle und Selbstwirksamkeit führt und dadurch weniger Gefahr besteht, dass Systementscheidungen blind übernommen werden.

Schließlich wächst die Erkenntnis, dass rein datengetriebenes maschinelles Lernen oft wenig effizient ist. Während Menschen bereits erworbenes Wissen und Kompetenzen nutzen und dadurch immer komplexere Dinge lernen können, wird beim maschinellen Lernen immer wieder alles von Neuem gelernt. Würde man Vorwissen in den Lernprozess einbeziehen können, könnte man Daten sparen, was wiederum zu weniger Aufwand für die Annotation sowie zu Einsparungen an Energie für Speicherung und Verarbeitung führen könnte. Zudem kann vorhandenes Wissen gezielt genutzt werden, um den Lernprozess zu steuern. Modelle, die vorhandenes Wissen berücksichtigen, sind weniger anfällig für unerwünschte Verzerrungen und robuster in Bezug auf Daten, die außerhalb der Datenverteilung im Training liegen.

Tiefe neuronale Netze haben das Forschungsgebiet Künstliche Intelligenz nach langen Jahren wieder ins öffentliche Interesse gebracht. Die zunehmende Digitialisierung und globale Vernetzung ermöglicht das Lernen aus großen Datenmengen. Für einen verantwortungsvollen Einsatz von KI-Methoden liefern die neuen Forschungsthemen der erklärbaren, interaktiven und hybriden KI die Chance, dass partnerschaftliche KI-Systeme entstehen, die menschliche Kompetenzen nicht beschneiden sondern erweitern und fördern.

Kollisionsrechtliche Fragen an die Nachvollziehbarkeit und Überprüfbarkeit von KI-Systemen

Kai v. Lewinski[1]

I. Einleitung

Künstliche Intelligenz ist neu,[2] spannend – und unbekannt. Sie ist uns auch noch unheimlich. Deshalb möchten Menschen, jedenfalls wenn KI in den Alltag Einzug hält, diesen neuen Mitbewohner und Mitbürger verstehen können. Hierauf richtet sich die (rechtspolitische) Forderung nach Nachvollziehbarkeit und Überprüfbarkeit von KI-Systemen. Ohne eine solche Nachvollziehbarkeit laufen wir Gefahr, dass die Ergebnisse nicht unseren Erwartungen entsprechen: KIs haben kein Konzept von richtig oder falsch, sondern nur von Zielen, was leicht zu sehr unerwarteten und ungewollten Lösungswegen führen kann.[3] Doch sind die Entscheidungen von KI-Systemen oft nicht im Nachhinein nachvollziehbar.[4] Die naturwissenschaftliche Aufbereitung dieser Frage steht indessen erst am Anfang.[5] Plastisch spricht man von der „Blackbox Algorithmus" und attestiert der algorithmischen Welt, sie gründe „ihr Fundament auf Arkan-Formeln".[6]

Wenn man sich die bestehenden Regelungen (dazu sogleich 1.–5.) anschaut, sieht man schnell, dass sie dieses Problem durchaus in Teilen und Ausschnitten adressieren, aber sicherlich nicht konsistent in den Griff bekommen. Künftigen, schon im Entwurf vorliegenden Regelungen (dann im Anschluss 6. u. 7.) wird dies sicherlich sehr viel besser gelingen. Bestehendes wie kommendes Recht aber haben einen blinden Fleck, insofern sie die internationale Dimension (s. u. III.) nicht (wirklich) in den Blick nehmen, was territorialen Regeln in der digitalen Welt viel von ihrer Wirksamkeit nimmt.

---

[1] Der *Verf.* dankt seinem Assistenten *Marvin Gülker* für die Unterstützung und Anregungen v. a. zu Teil II. dieses Texts.

[2] Neu jedenfalls auf der Agenda des Normgebers. Informatisch neu ist KI mitnichten. Die ersten Erkenntnisse gehen zurück bis auf die 40er Jahre, der englische Ausdruck Artificial Intelligence stammt wohl aus einem Papier von 1956 (s. *Russell/Norvig*, Artificial Intelligence: A Modern Approach, 4. Aufl. 2020, 17 f.).

[3] *Russell/Norvig*, Artificial Intelligence: A Modern Approach, 4. Aufl. 2020, 5; anschaulich zum „Clever Hans Effect" *Lapuschkin u. a.*, Nature Communications 2019, 1096.

[4] COM(2020) 65 final, 14; zu den schwerverständlichen „hidden layers" *Russel/Norvig* (Fn. 3), 759.

[5] Zum aktuellen Stand von „Explainable AI" *Höhne*, DuD 2021, 453.

[6] *Martini*, JZ 2017, 1017 (1018).

II. Bestehende und kommende Regelungen

1. Verbraucherschutzrecht

Das Verbraucherschutzrecht kennt (noch) keine expliziten Regelungen zur Künstlichen Intelligenz. Allerdings unterfallen auch „smarte" Produkte mit KI-Komponenten dem Produkthaftungsrecht, sodass bei Verletzung geschützter Rechtsgüter ein Anspruch aus § 1 Abs. 1 ProdHaftG möglich ist,[7] der verschuldensunabhängig ist und insoweit das Nachvollziehbarkeitsproblem umgeht. Streitig ist in diesem Zusammenhang allerdings, ob reine Software überhaupt den Produktbegriff des § 2 ProdHaftG erfüllt.[8] Und auch die Nachvollziehbarkeitsproblematik taucht auf den zweiten Blick dann doch auf: nämlich bei der Frage, ob die Ersatzpflicht des Herstellers gem. § 1 Abs. 2 Nr. 5 ProdHaftG wegen Unvorhersehbarkeit nach dem Stand der Wissenschaft und Technik ausgeschlossen ist.[9] Ferner sind Ansprüche aus der Produzentenhaftung nach § 823 Abs. 1 BGB zumindest denkbar. Der im Deliktsrecht notwendige Nachweis der Kausalität stellt angesichts des geschilderten Nachvollziehbarkeitsproblems die Rechtsordnung allerdings auch hier vor Probleme.[10] Zur Lösung hat man die Einführung einer Rechtspersönlichkeit für Künstliche Intelligenzen diskutiert, was mit Recht unter Verweis auf die Konzeption des Deliktsrechts als an menschlichem Verhalten orientiert abgelehnt wird.[11] Eine zufriedenstellende Lösung ist für das Deliktsrecht bisher nicht gefunden worden, ausdrückliche Regelungen bestehen nicht.

Nun mag man einwenden, dass sich schon aus dem allgemeinen Vertragsrecht ergebe, dass die wesentlichen Eigenschaften eines Vertragsgegenstands, ganz unabhängig von KI, der Vertragsgegenseite bekannt sein müssen.[12] „Fehlerhafte" KI-Entscheidungen können sich als Mängel darstellen.[13] Dafür muss man allerdings überhaupt erst einmal einen Vertrag geschlossen haben; viele KI-Systeme im Verbraucherkontext sind aber in der vorvertraglichen Phase eingesetzt, so dass hier eine Lücke besteht.

So oder so scheint die zivilrechtliche Diskussion vor allem um die Haftungsfrage zu kreisen, adressiert die mangelnde Nachvollziehbarkeit deshalb nicht als

---

[7] *Eisenberg*, InTeR 2021, 17 (18).

[8] Zum Meinungsstand *Joggerst/Wendt*, InTeR 2021, 13 (13 f.); ausf. *Wagner*, in: MüKo BGB, 8. Aufl. 2020, § 2 ProdHaftG Rn. 21 ff.

[9] *Schaub*, JZ 2017, 342 (343).

[10] *Schaub*, in: Taeger, Rechtsfragen digitaler Transformationen, 2018, 439 (440 f.).

[11] *Denga*, CR 2018, 69 (77) mit Verweis auf den Perserkönig *Xerxes*, der *Herodot* zufolge das Meer auspeitschen ließ.

[12] *Micklitz/Namysłowska/Jabłonowska*, in: Ebers/Heinze/Krügel/Steinrötter, Künstliche Intelligenz und Robotik, 2020, § 6 Rn. 106. – Für Anwälte wird als Ausprägung der berufsrechtlichen Sorgfalt gegenüber dem Mandanten Transparenz hinsichtlich des Einsatzes von KI-Systemen und v.a. deren Schwächen („Unschärfen") verlangt (*Fries*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 15.1 Rn. 25).

[13] *Schaub*, JZ 2017, 342 (343).

eigenständiges Rechtsproblem. Wo das Ergebnis der KI-„Entscheidung" nicht erkennbar rechtswidrig ist, wird ein Problem nicht gesehen.

2. Antidiskriminierungsrecht

KIs haben Einzug gehalten in die Arbeitswelt, und mit ihnen die typischen Probleme derartiger Systeme.[14] Grund hierfür ist zwar nicht, dass KI-Programme „wissentlich" bestimmte gesellschaftliche Gruppen benachteiligen „wollten" – dazu sind heutige KI-Programme mangels des dafür notwendigen Bewusstseins überhaupt nicht in der Lage,[15] und derartige „starke" KI-Systeme bleiben 2023 weiterhin bestenfalls Science Fiction.[16] Vielmehr verbirgt sich hinter einer derartigen vermeintlichen Diskriminierung „durch die Software" die Erkennung und Fortschreibung bestehender Diskriminierungen. – KI schafft nicht per se Diskriminierung, sondern andere. Es besteht aber die Gefahr, dass bestehende gesellschaftliche Vorurteile durch den scheinbar objektiven Algorithmus festgeschrieben werden.[17]

In der Arbeitswelt gilt das Antidiskriminierungsrecht. Und zuweilen finden sich solche Korrelationen dann eben auch mit Bezug auf die laut § 1 AGG verbotenen Kriterien, wodurch namentlich die Gefahr einer mittelbaren Diskriminierung im Sinne des § 3 Abs. 2 AGG begründet wird.[18] Geht man davon aus, dass die Bewertung durch einen KI-Algorithmus eine „Behandlung" im Sinne des § 3 Abs. 1 AGG darstellen kann oder ihr eine solche doch zumindest nachfolgt,[19] wird der Arbeitgeber sich jedenfalls im Rahmen des materiellen Schadensersatzanspruchs aus § 15 Abs. 1 AGG nicht damit herausreden können, das „autonome" Handeln des von ihm eingesetzten KI-Algorithmus sei ihm nicht zurechenbar.[20]

Nachvollziehbarkeit und Überprüfbarkeit als zentrale Problematik des Einsatzes von KI-Systemen adressiert das Antidiskriminierungsrecht demgegenüber jedoch nicht. Die stattdessen vorgesehene Beweislasterleichterung des § 22 AGG (in Umsetzung von Art. 10 RL 2000/78/EG) greift erst nach Darlegung hinreichender Indizien für eine Diskriminierung ein. Weil die Funktionsweise der oft

---

[14] So wird von Amazon berichtet, dass das Unternehmen einen selbstlernenden Algorithmus zur Beurteilung von Bewerbern wieder abschalten musste, weil er systematisch Frauen benachteiligte (F.A.Z. vom 8.3.2021, S. 18).

[15] Heutige KI-Systeme erkennen Korrelationen, die auch absurd ausfallen können: So soll es eine 99 %ige Korrelation zwischen der Scheidungsrate im US-Bundesstaat Maine und dem Pro-Kopf-Verzehr von Margarine geben (*Dzida/Groh*, NJW 2018, 1917 [1918]).

[16] *Russel/Norvig* (Fn. 3), 32 f.

[17] So für den in Österreich bei der Arbeitslosenvermittlung eingesetzten AMS-Algorithmus *Wagner u. a.*, juridikum 2/2020, 191 (199).

[18] *v. Lewinski/de Barros Fritz*, NZA 2018, 620 (622).

[19] *v. Lewinski/de Barros Fritz*, NZA 2018, 620 (621).

[20] *Dzida/Groh*, NJW 2018, 1917 (1920); *v. Lewinski/de Barros Fritz*, NZA 2018, 620 (623). Beim immateriellen Schadensersatzanspruchs aus § 15 Abs. 2 AGG ist das nicht so klar (s. *v. Lewinski/de Barros Fritz*, a. a. O., 624).

eingekauften KI-Programme meist Geschäftsgeheimnis und für Laien – wie wohl auch oft für Experten – nicht nachvollziehbar ist, wird schon das Darlegen dieser Indizien ein schwieriges Unterfangen sein.[21]

3. Datenschutzrecht

Das Datenschutzrecht kommt konzeptionell von der guten alten Datenverarbeitung her (auch wenn es inhaltlich nicht um Daten, sondern Informationen geht). Algorithmen (s. aber u. c)) und auch KI werden vom Datenschutzrecht nicht unmittelbar adressiert; es fällt ihm schwer, KI als eine nur indirekt auf Daten aufsetzende Technik zu erfassen. Auch gibt es keine konzeptionelle Verbindung zwischen Datenschutzgesetzgebung und KI-Regulierung.[22]

a) Transparenz

Anders als das Antidiskriminierungsrecht kennt das Datenschutzrecht weitgehende Informierungs- und Transparenzpflichten. Doch laufen diese bei Blackbox-Systemen weitgehend ins Leere – wenn auch die verantwortliche Stelle nicht weiß, warum das System etwas macht, kann auch nichts beauskunftet werden.[23]

b) Technisch-organisatorische Maßnahmen und Datensicherheit

Technisch-organisatorische Vorgaben des Datenschutzrechts richten sich auf datenschutzrechtliche Fragen. Ziel der Regelungen der technisch-organisatorischen Sicherheit in der DSGVO ist es, die Verletzung des Schutzes personenbezogener Daten technikneutral zu verhindern, weshalb sich der KI-Begriff in den datensicherheitsrechtlichen Regelungen der DSGVO nicht findet.[24] Das gilt insbesondere auch für das „Privacy by Design" des Art. 25 Abs. 1 DSGVO. Selbst wenn man darin ein verallgemeinerungsfähiges Prinzip erblicken möchte,[25] adressiert die eher vage Formulierung der Norm nicht alle informationstechnischen Wünsche und Problemlagen. Zwar erscheint es durchaus sinnvoll, die Berücksichtigung gesamtgesellschaftlicher Vorstellungen schon in der Programmierung von Software gesetzlich vorzuschreiben,[26] doch ist das keine Frage allein des Datenschutzrechts. Der Datenschutz ist kein Alleskleber für jede technische Herausforderung.

---

[21] *Dzida/Groh*, NJW 2018, 1917 (1922).

[22] *Schallbruch*, DuD 2021, 438 (443) mit dem zusätzlichen Hinweis, dass auch IT-Sicherheitsrecht und KI-Regulierung ebenfalls nur oberflächlich (vgl. Art. 42 Abs. 2 EU-KI-VO-KommE) verknüpft wären.

[23] *Kroschwald*, DuD 2021, 522 (525).

[24] *Kipker/Müller*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, 2020, Kap. 8.6 Rn. 4 f.

[25] So *Eisenberg*, InTeR 2021, 17, 19 ff., der Art. 25 I DSGVO in die deliktischen Verkehrssicherungspflichten hineinlesen will.

[26] *Martini*, JZ 2017, 1017 (1019).

Allerdings hält auch das genuine IT-Sicherheitsrecht keine eigenen Regelungen für Künstliche Intelligenzen bereit. Soweit § 8a BSIG die Betreiber Kritischer Infrastrukturen in die Pflicht nimmt, macht die Norm keine spezifischen Vorgaben für KI, sondern erschöpft sich ähnlich wie Art. 32 DSGVO im unbestimmten Rechtsbegriff vom „Stand der Technik".[27] § 8a Abs. 4 BSIG ermöglicht dem BSI zwar die Überprüfung Kritischer Infrastrukturen bis hin zum Betreten der Geschäfts- und Betriebsräume und der Einsicht in Unterlagen, doch hilft auch das wenig bei der Nachvollziehung der Entscheidungswege einer KI. Das bereichsspezifische IT-Sicherheitsrecht, das man als unsystematischen „Regelungszoo" bezeichnet hat,[28] enthält ebenfalls keine spezifischen Vorgaben für KI.[29]

c) Automatisierte Einzelentscheidung

Allerdings gibt es innerhalb des Datenschutz-Rechtsakts in Gestalt der „Automatisierten Einzelentscheidungen (einschließlich Profiling)" in Art. 22 DSGVO eine Regelung für algorithmische Entscheidungssysteme. Sie passt zwar fast eher auf lochkartengestützte Entscheidungsbäume als auf neuronale Netze, gleichwohl aber sind die Regeln auch auf KI-Systeme anwendbar.[30] Die Norm verlangt eine gewisse Komplexität der Entscheidungsfindung, da sonst die besonderen Voraussetzungen des Art. 22 auf triviale Angelegenheiten wie das Abheben von Geld am Geldautomaten Anwendung fänden.[31] Zumindest dann, wenn KI auf der Grundlage maschinellen Lernens zum Einsatz kommt, wird angesichts der Blackbox-Problematik regelmäßig die erforderliche Komplexität vorliegen.[32] Unabhängig von der Frage, was genau man eigentlich unter Profiling selbst unter der Zuhilfenahme der Legaldefinition in Art. 4 Nr. 4 DSGVO verstehen soll,[33] regelt die Norm aber jedenfalls nicht die eigentliche technische Anwendung der KI, sondern nur die nachgelagerte Entscheidungsfindung.[34] Geht die endgültige Entscheidung nicht allein von der Maschine aus, beruht die Entscheidung nicht „ausschließlich" auf dem Profiling und dann greift Art. 22 DSGVO insgesamt nicht ein. Auch wenn man dieses Kriterium nicht schon durch einen menschlichen „Abstempler" umgehen kann,[35] schließt es doch eine erhebliche Menge von KI-basierten Systemen von der Anwendung des Art. 22 DSGVO aus.

---

[27] Vgl. *Kipker/Müller*, in: Kaulartz/Braegelmann (Fn. 24), Kap. 8.6 Rn. 6; zum Inhalt dieses Begriffs ausführlich *Ekrot/Fischer/Müller*, in: Kipker, Cybersecurity, 2020, Kap. 3 Rn. 1 ff.

[28] *Ritter*, zit. nach *Blach*, GRUR 2021, 1038 (1039).

[29] Vgl. *Blach*, GRUR 2021, 1038 (1039).

[30] Verweis auf *Spielkamp*, in diesem Band.

[31] *v. Lewinski*, in: Wolff/Brink, Datenschutzrecht, 2. Aufl. 2022, Art. 22 DSGVO Rn. 13.

[32] *v. Lewinski*, in: Wolff/Brink, Datenschutzrecht, 2. Aufl. 2022, Art. 22 DSGVO Rn. 13.

[33] Dazu *v. Lewinski*, in: Wolff/Brink, Datenschutzrecht, 2. Aufl. 2022, Art. 22 DSGVO Rn. 7.

[34] *Lorentz*, Profiling, 2020, 78 f. und 156.

[35] Der Mensch muss am Ende eingeschaltet werden, kompetent sein und eigenen Entscheidungsspielraum besitzen (*v. Lewinski*, in: Wolff/Brink, Datenschutzrecht, 2. Aufl. 2022, Art. 22 DSGVO Rn. 23 ff.).

Auch statuiert das Datenschutzrecht ausdrückliche Unterrichtungspflichten bei Automatisierter Einzelentscheidung i. S. d. Art. 22 DSGVO hinsichtlich „aussagekräftiger Informationen über die involvierte Logik" (Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g DSGVO), ebenso einen entsprechenden Auskunftsanspruch (Art. 15 Abs. 1 lit. h DSGVO). Mit Blick auf die dargestellten Diskriminierungspotentiale (s. o. 2.) soll dieser Auskunftsanspruch Informationen über die statistische Verteilung des Trainingsmodells umfassen.[36] Ausdrücklich normiert ist dies indes nicht, sodass über diesen Punkt Streit besteht.[37] Jedenfalls aber besteht ein Widerspruchsrecht (vgl. die ausdrückliche Erwähnung von „Profiling" in Art. 21 Abs. 1 S. 1 Hs. 2 DSGVO).

Die Trainingsphase, in der das Potential zu Diskriminierungen gelegt wird, regelt Art. 22 DSGVO aber überhaupt nicht, sodass man lediglich über die Anwendung bzw. Vorwirkung der allgemeinen Regelungen des Datenschutzrechts nachdenken kann.[38] Nicht mit der Systematik des Datenschutzrechts und der Begrenzung seines Anwendungsbereichs auf die konkrete Verarbeitung personenbezogener Daten vereinbar ist es, den Grundsatz der Datenrichtigkeit aus Art. 5 Abs. 1 lit. d DSGVO in die Trainingsphase vorzuverlegen.[39] Hier wird versucht, dem Datenschutzrecht fremde Erwägungen (die spezifischen Risiken von KI) einzubinden.

Im Schrifttum vorgeschlagen wird ferner als kompensatorische „angemessene Maßnahmen" für rechtlich oder ähnlich beeinträchtigende Entscheidung u. a. die Erklärbarkeit (Explainability) des Algorithmus bzw. der wesentlichen Parameter der Entscheidungsfindung.[40] Derartige Informationen dürften allerdings regelmäßig ein schützenswertes Geschäftsgeheimnis bilden.[41]

4. Wettbewerbsrecht

Das Wettbewerbsrecht enthält keine ausdrücklichen Regelungen zu Künstlicher Intelligenz. Allerdings ist es sehr wohl denkbar, dass ein selbstlernender Algorithmus Entscheidungen trifft, die dann eine Wettbewerbsrechtsverletzung darstellen. So ist etwa darüber diskutiert worden, ob die Aufnahme von Konkurrenzprodukten in eine Trefferliste durch einen selbstlernenden Algorithmus eine möglicherweise unzulässige vergleichende Werbung nach § 6 Abs. 1 UWG darstellt.[42] Denkbar sind auch Fälle der Irreführung nach § 5 Abs. 1 UWG einschließlich irreführender Unterlassungen nach § 5a UWG, wobei man wohl beim Einsatz selbstlernender Algorithmen eher von einem aktiven Tun als von einem

---

[36] *Hacker*, NJW 2020, 2142 (2144).

[37] Ausführlich *Dimitrova*, EDPL 2020, 211 (212 ff.).

[38] Was sich bei synthetischen Daten schwierig gestaltet (s. *Raji*, DuD 2021, 303 [306 ff.]).

[39] So aber *Stevens*, CR 2020, 73 (74).

[40] *Hacker*, NJW 2020, 2142 (2144); *Sesing*, MMR 2021, 288 (292 f.); vorsichtiger *Martini*, JZ 2017, 1017 (1020).

[41] So zur SCHUFA-Scoreformel noch vor Inkrafttreten der DSGVO *BGH*, MMR 2014, 489 (491); zur Fortgeltung der Erwägung *v. Lewinski/Pohl*, ZD 2018, 17 (22).

[42] *Ohly*, WRP 2018, 131 (138); *Schaub* (Fn. 10), 446. – Der EuGH (Urt. v. 23.3.2010 – C-236/08 u. a., GRUR 2010, 445 Rn. 71) hat diese Frage beim Keyword Advertising offengelassen.

Unterlassen ausgehen muss.[43] Klare gesetzliche Maßstäbe fehlen, sodass es auf die allgemeinen Maßstäbe des Wettbewerbsrechts ankommt, die auf Künstliche Intelligenzen nicht recht passen. Jedenfalls stellt das Wettbewerbsrecht vor allem auf die Wahrnehmung des verständigen Durchschnittsverbrauchers (§ 3 Abs. 4 S. 1 UWG; sog. „europäisches Verbraucherleitbild")[44] ab. Dieser wird regelmäßig mit dem Nachvollziehen der konkret eingesetzten Technik überfordert sein. Hier mag man an Informationspflichten[45] denken, wenngleich es für die Frage, wie die von einer KI ausgegebenen Ergebnisse auf den Durchschnittsverbraucher *wirken*, auf ihre innere Funktionsweise gar nicht ankommt. Aus dem Wettbewerbsrecht können daher kaum Erkenntnisse zur Nachvollziehbarkeit von KI gewonnen werden.

5. Geschäftsgeheimnisschutz

Bei Algorithmen und KI kann es sich um Geschäftsgeheimnisse i. S. d. GeschGehG (und früher der §§ 17, 18 UWG a. F.) handeln[46] – selbst wenn der Geheimnisherr nicht recht weiß, was da in seiner Blackbox so vor sich geht. Ein entsprechendes wirtschaftliches Interesse besteht jedenfalls.

So hält das Unternehmen OpenAI seine KI „GPT-3" strikt unter Verschluss und gewährt Zugriff nur über eine Online-Schnittstelle.[47] Die „Lizenzierung" erfolgt dann vertragsrechtlich.[48]

Dieser Geschäftsgeheimnisschutz kann gegenüber der Aufdeckung von geheimgehaltenen Algorithmen geltend gemacht werden. Ohne hier auf Einzelheiten einzugehen zeigt sich, dass das GeschGehG und der Unternehmensgeheimnisschutz insgesamt der Nachvollziehbarkeit und Überprüfbarkeit nicht förderlich sind.

6. EU-KI-VO

Die Europäische Kommission hat im April 2021 einen Entwurf für eine EU-KI-VO (im folgenden EU-KI-VO-KommE) vorgelegt.[49] Als erstem spezifisch und

---

[43] *Schaub* (Fn. 10), 446.

[44] *Micklitz/Namysłowska/Jabłonowska*, in: Ebers/Heinze/Krügel/Steinrötter, Künstliche Intelligenz und Robotik, 2020, § 6 Rn. 34 ff., 40 ff.

[45] *Micklitz/Namysłowska/Jabłonowska*, in: Ebers/Heinze/Krügel/Steinrötter, Künstliche Intelligenz und Robotik, 2020, § 6 Rn. 101 ff.

[46] *Apel/Kaulartz*, RDi 2020, 24 (29 f.); *Bischof/Intveen*, ITRB 2019, 134 (136); *Söbbing*, MMR 2021, 111 (115 f.).

[47] *Stieler*, c't 21/2021, 124 (125).

[48] Für GPT-3 *Stieler*, c't 21/2021, 124 (126); konkrete Formulierungsvorschläge bei *Apel/Kaulartz*, RDi 2020, 24 (32 f.).

[49] *EU-Kommission*, 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, COM(2021) 206 final v. 21.4.2021.

bereichsübergreifend auf KI zugeschnittenen Rechtsakt überhaupt kommt dem Vorschlag eine hohe symbolische Bedeutung zu.[50]

Der Entwurf folgt – durchaus anders als im Datenschutzrecht[51] – einem risikobasierten Ansatz, wie ihn etwa auch die Datenethikkommission vorgeschlagen hatte.[52] Er unterscheidet[53] zwischen wegen eines unannehmbaren Risikos verbotenen KI-Systemen (Tit. II EU-KI-VO-KommE), Hochrisikosystemen (Tit. III EU-KI-VO-KommE; s. dazu u. d)) und Systemen mit geringem Risiko (s. dazu u. b)); sonstige, nur minimal riskante Systeme werden nicht (unmittelbar) geregelt (s. a)).

Den notorisch unscharfen Begriff KI[54] definiert die Verordnung gem. Art. 3 Nr. 1 EU-KI-VO-KommE über die Verwendung bestimmter, im Anhang I zur Verordnung aufgeführter Techniken und damit anders als erwartet[55] weder über symbolische noch über subsymbolische Systeme.[56] Im Ergebnis dürften mehr als bloß selbstlernende Systeme vom Verordnungsentwurf erfasst sein.[57]

a) Keine zusätzlichen Pflichten für einfache Systeme

Die niedrigste von der KI-Verordnung erfasste Stufe bilden KI-Systeme, die nach Ansicht der Kommission ein minimales Risiko darstellen.[58] An sie stellt der Vorschlag keine besonderen inhaltlichen Anforderungen (vgl. ErwGr. 14 EU-KI-VO-KommE), sodass eine Berücksichtigung freiwillig bleibt.[59] Dem Wortlaut des Art. 25 Abs. 1 EU-KI-VO-KommE nach müssen allerdings alle nicht in der Union ansässigen Anbieter jeglicher KI-Systeme einen Bevollmächtigten bestellen. Da sich dessen Aufgaben gem. Art. 25 Abs. 2 EU-KI-VO-KommE aber nur auf Hochrisiko-KI-Systeme beziehen und die Norm sich im Kontext mit Regelungen zu Hochrisiko-KI-Systemen und unterhalb der Abschnittsüberschrift „Pflichten der Anbieter und Nutzer von Hochrisiko-KI-Systemen und anderer Beteiligter" befindet, ist „Anbieter" aber wohl so zu verstehen, dass der Begriff im Einklang mit den umstehenden Bestimmungen „Anbieter von Hochrisiko-KI-Systemen" bedeutet. Ein eigenes Regelungsregime für einfache Systeme errichtet der Entwurf daher nicht.

---

[50] *Unger*, ZRP 2020, 234 (235); vgl. auch *Ebert/Spiecker*, NVwZ 2021, 1188 (1188): „weltweit einzigartiger Aufschlag".

[51] A. A. aber *Bomhard/Merkle*, RDi 2021, 276 (277 [Rn. 3]): „präventives Verbotsgesetz".

[52] Datenethikkommission, Gutachten (2019), 173 ff.

[53] Die Datenethikkommission, Gutachten (2019), 177 hatte sogar eine Kritikalitätspyramide mit fünf Stufen vorgeschlagen.

[54] Zur Begriffsproblematik *Herberger*, NJW 2018, 2825 (2825 ff.).

[55] *Unger*, ZRP 2020, 234 (235).

[56] *Schallbruch*, DuD 2021, 438 (441).

[57] *Ebert/Spiecker*, NVwZ 2021, 1188 (1188 f.); *Spindler*, CR 2021, 361 (363).

[58] COM(2021) 602 final, Ziff. 5.2.2.

[59] *Engelmann/Brunotte/Lütkens*, RDi 2021, 317 (321 [Rn. 25]); *Ebert/Spiecker*, NVwZ 2021, 1188 (1188).

b) Transparenz- und Kennzeichnungsanforderungen für bestimmte Systeme

Für KI-Systeme mit einem bestimmten Risiko bestehen Transparenz- (Tit. IV EU-KI-VO-KommE) und Kennzeichnungspflichten.[60] Systeme, die mit Menschen interagieren oder (Medien-)Inhalte generieren, müssen so gestaltet sein, dass ein Mensch erkennen kann, dass er es mit einem KI-System zu tun hat (ErwGr. 70 und Art. 52 EU-KI-VO-KommE). Dies gilt ausdrücklich auch für Systeme, die Gefühlsregungen von Menschen erkennen können (ErwGr. 70 S. 4 EU-KI-VO-KommE).

Solche Kennzeichnungspflichten waren schon früh gefordert worden.[61] Doch wird auch kritisiert, dass die Kennzeichnungspflichten ins Leere liefen, weil in Ermangelung von ausdrücklichen Betroffenenrechten,[62] einer Beweislastumkehr o. ä. eine menschliche Einflussnahmemöglichkeit gar nicht bestehe.[63]

Gerade in dieser Ausgestaltung kann man hierin auch übertriebenes, aber konkret folgenloses Warnen in Konstellationen technischen und sozialen Wandels sehen – allgemein wird das als „Red Flag Traffic Law" bezeichnet, was auf die Zeit der Erfindung des Automobils zurückgeht, als dem Fahrzeug ein Mensch mit roter Warnflagge vorangehen musste.

c) Anforderungen an Maschinen mit KI (Maschinen-VO-E)

Zeitgleich mit dem Regulierungsvorschlag für KI-Systeme hat die Europäische Kommission auch einen Vorschlag für eine Maschinenverordnung[64] (EU-MaschinenVO-KommE), der die Maschinenrichtlinie 2006/42/EG ablösen soll. Sie koordiniert ihre Regelungen für Geräte, in die KI-Komponenten eingebaut sind, mit der EU-KI-VO (ErwGr. 19 und Art. 9 EU-MaschinenVO-KommE).

In Transparenz- und Nachvollziehbarkeitshinsicht ist das CE-Kennzeichen für Importe (ErwGr. 32–35 EU-MaschinenVO-KommE) als das „sichtbare" Zeichen der Konformität (ErwGr. 47 EU-Maschinen-VO-KommE) zu erwähnen. Auf den ersten Blick scheint es paradox, einem algorithmischen System ein körperliches Zeichen umhängen zu wollen; durch den Kontext der MaschinenVO, die ja immer ein anfassbares Gerät, auf das man einen Aufkleber kleben kann, voraussetzt, ist die Kennzeichnung aber doch sinnvoll.

Ausdrücklich geregelt ist zudem, dass auf Messen und Ausstellungen Maschinen von den Vorgaben der MaschinenVO abweichen dürfen, wenn das Publikum entsprechend informiert ist (ErwGr. 20 und Art. 4 EU-MaschinenVO-KommE).

---

[60] *Schallbruch*, DuD 2021, 438 (442).

[61] *Martini*, JZ 2017, 1017 (1020).

[62] „Betroffene" spielen keine Rolle: Sie werden nicht einmal legaldefiniert, und sie haben – anders als im Datenschutzrecht – auch keine *Betroffenen*rechte (*Bomhard/Merkle*, RDi 2021, 276 [283, Rn. 44]); *Ebert/Spiecker*, NVwZ 2021, 1188 (1193).

[63] *Ebert/Spiecker*, NVwZ 2021, 1188 (1191).

[64] *Europäische Kommission*, Vorschlag für eine Verordnung des europäischen Parlaments und des Rates über Maschinenprodukte, COM(2021) 202 final v. 21.4.2021.

d) Erhöhte Anforderungen an Hochrisikosysteme

Tit. III Abschn. 2 EU-KI-VO-KommE widmet sich Hochrisikosystemen. Anders als bei der DSGVO (s. o. 3.) wird stärker direkt bei der Entwicklung der KI angesetzt und nicht bloß beim einzelnen Verarbeitungsschritt.[65]

aa) Nachvollziehbarkeit des Einsatzes

Für Hochrisikosysteme besteht eine ganze Reihe von Anforderungen, um deren Einsatz nachvollziehbar zu machen.

Es besteht eine Registrierungspflicht für Hochrisiko-KI-Systeme (Tit. VII EU-KI-VO-KommE). Technische Dokumentation (Art. 11, Art. 18 EU-KI-VO-KommE; diese Regelungen haben durchaus Verwandtschaft mit Art. 10 EU-MaschinenVO-KommE) muss vorhanden sein, ebenso Aufzeichnungen über den Betrieb (Art. 12, Art. 20 EU-KI-VO-KommE). Vorgeschrieben sind weiter ein Qualitätsmanagementsystem (Art. 17 EU-KI-VO-KommE) und für Anbieter von Hochrisikosystemen mit Meldepflicht bei Vorfällen eine Post-Market-Beobachtungspflicht (ErwGr. 78 und Art. 61 EU-KI-VO-KommE).

Ausdrücklich angeordnet ist für Hochrisikosysteme auch eine Nutzerverantwortlichkeit hinsichtlich des Einsatzes (ErwGr. 58 und Art. 29 EU-KI-VO-KommE), was jedenfalls die Verantwortlichkeit einer rechtlichen Person für Hochrisikosysteme (ErwGr. 53 EU-KI-VO-KommE) nachvollziehbar macht.[66]

Das Diskriminierungsproblem (vgl. o. 2.) soll auf der Trainingsdatenebene angegangen werden, die Normierung der Anforderungen an die Trainingsdaten in Art. 10 EU-KI-VO-KommE regelt eigentlich eine Selbstverständlichkeit.[67] Das Lernen im laufenden Betrieb beim Nutzer ist allerdings nicht geregelt.[68] Außerdem soll sich der nach Art. 14 EU-KI-VO-KommE notwendige menschliche Aufseher nach dessen Abs. 4 lit. b des „Automatisierungsbias" bewusst sein, d. h. nicht einfach Empfehlungen der KI als objektiv richtig annehmen und übernehmen.[69] Spezifisch gegen das Blackbox-Problem ist nach Meinung *Spindlers* der Einsatz von Logging Devices nach Art. 12 Abs. 1 EU-KI-VO-KommE gerichtet.[70] Die Norm schreibt allerdings nur vor, Vorgänge und Ereignisse zu protokollieren; die Art und Weise, wie eine Entscheidung zustandegekommen ist, kann man daraus nicht ersehen. Immerhin hat man damit aber Rahmeninformationen dafür an der Hand, welche Eingabe eine als unrichtig empfundene „Entscheidung" hervorgerufen hat.

---

[65] *Ebert/Spiecker*, NVwZ 2021, 1188 (1188); Forderung schon bei *Martini*, JZ 2017, 1017 (1019).
[66] Zugleich ist dies eine (implizite) Absage an die „elektronische Rechtsperson".
[67] *Spindler*, CR 2021, 361 (367).
[68] *Spindler*, CR 2021, 361 (368).
[69] *Spindler*, CR 2021, 361 (367).
[70] *Spindler*, CR 2021, 361 (367).

bb) Nutzer- und Betroffeneninformation

Nutzer sollen angesichts der Undurchsichtigkeit („Opacity") Dokumentation und Anleitungen erhalten, um den Output von Hochriskio-KI-Systemen beurteilen zu können (ErwGr. 47 EU-KI-VO-KommE). Hochrisikosysteme sollen so konstruiert sein (Design-Vorgabe), dass Menschen ihr Funktionieren überwachen können (ErwGr. 58 und Art. 14 EU-KI-VO-KommE). Die Transparenz und Nutzerinformation (Art. 13 EU-KI-VO-KommE) sollen den Maßstäben (verständlicher) Erklärbarkeit (Explainability oder auch Explainable AI [XAI]) entsprechen.[71]

Mit Blick auf den Einsatz sollen auch die jeweiligen Fähigkeiten der Nutzer berücksichtigt werden (Art. 9 Abs. 4 UAbs. 3 EU-KI-VO-KommE). Anbieter von Hochrisiko-KI müssen deshalb verständliche Aufzeichnungen und technische Dokumentation be(reit)halten (ErwGr. 46 EU-KI-VO-KommE). In Anbetracht der technischen Entwicklung auf diesem Gebiet allgemein und der Dynamik von KI-Systemen speziell ist die Maßgabe, dass die technische Dokumentation aktuell zu halten sei (ErwGr. 46 S. 4 EU-KI-VO-KommE), fast schon putzig. Ob und unter welchen Bedingungen die hierbei relevanten technischen Standards entwickelt werden, steht in den Sternen.[72]

7. Digital Markets Act und Digital Service Act

Ebenfalls im europäischen Gesetzgebungsverfahren befanden bzw. befinden sich der Digital Markets Act[73] und der Digital Services Act[74]. Der Digital Markets Act definiert in Art. 3 DMA-KommE gewisse marktstarke Plattformen als sogenannte „Gatekeeper", die umfassenden Pflichten nach Kap. III des Vorschlags unterliegen. So enthält Art. 3 lit. a DMA-KommE ein Verbot der Zusammenführung personenbezogener Daten über Dienstgrenzen hinweg, sofern keine Einwilligung nach Maßgabe der DSGVO vorliegt. Diese Regelung wird den Einsatz von KI durch den Gatekeeper beeinträchtigen, da ihm vorbehaltlich der Einwilligungsmöglichkeit die Möglichkeit genommen wird, KI-Algorithmen zur Erkennung von Nutzerpräferenzen auf den gesamten denkbaren Datenbestand zu einem Nutzer anzusetzen. Bekanntlich hängt aber die Qualität von KI von der Menge der verwendeten Daten ab. Ähnliches gilt für die Regelung des Art. 6 Abs. 1 lit. a DMA-KommE, die die Verwendung prinzipiell vorhandener, lediglich nicht öffentlicher Daten zum Schutz der Mitbewerber untersagt. Die in Art. 6 Abs. 1 lit. d DMA-

---

[71] Dazu *Hoffmann/Hevekordes*, DuD 2021, 609 (611).

[72] *Spindler*, CR 2021, 361 (367).

[73] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über bestreitbare und faire Märkte im digitalen Sektor (Gesetz über digitale Märkte), COM(2020) 842 final.

[74] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie 2000/31/EG, COM(2020) 825 final.

KommE normierte Pflicht, Dienstleistungs- und Produktrankings „anhand fairer und diskriminierungsfreier Bedingungen vorzunehmen", verlangt beim Einsatz von KI-Systemen wegen deren Anfälligkeit für diskriminierende Ergebnisse (s. o. 2.) besondere Sorgfalt. Demgegenüber dürften die Interoperabilitätspflichten nach Art. 6 Abs. 1 lit. f–g DMA-KommE nicht so weit gehen, Zugriff auf etwaige eingesetzte KIs zu erzwingen, da dies für eine erfolgreiche Interoperabilität nicht notwendig ist.

Der undurchsichtige Einsatz Künstlicher Intelligenz auf den Meinungsaustauschplattformen der großen Informationsintermediäre hat Anlass zur Sorge um den demokratischen Diskurs gegeben.[75] Dem will die Kommission mit dem Digital Services Act entgegentreten, der zwar ähnlich wie der Digital Markets Act in seiner derzeitigen Fassung nicht ausdrücklich von der Regulierung künstlicher Intelligenz spricht. Anders als beim Digital Markets Act zielen aber einige Regelungen auf Probleme, die spezifisch mit dem Einsatz von künstlicher Intelligenz zusammenhängen. Die nach Plattformgröße abgestuften Pflichten in Kap. III DSA-KommE verpflichten etwa in Art. 12 Abs. 1 DSA-KommE zur Offenlegung der Tatsache, dass Maßnahmen „algorithmischer Entscheidungsfindung" bei der Beurteilung von Nutzerinhalten eingesetzt werden. Das in Art. 15 Abs. 1 statuierte Erfordernis einer „klaren und spezifischen Begründung" für Sperrungen wird in Art. 15 Abs. 2 lit. c DSA-KommE ausdrücklich um die Pflicht ergänzt, den Einsatz „automatischer Mittel zur Entscheidungsfindung" offenzulegen – die Zusammenschau beider Normen lässt allerdings erkennen, dass nicht offenzulegen ist, wie das „automatisierte Mittel" zu seiner Entscheidung gekommen ist. Insoweit scheint die Kommission sich dem Blackbox-Problem zu beugen. Bemerkenswert ist das deshalb, weil Art. 17 Abs. 5 DSA-KommE, der verlangt, dass Entscheidungen „nicht allein mit automatisierten Mitteln getroffen werden" dürfen, gem. Art. 16 DSA-KommE für kleine Unternehmen nicht gilt, diesen also den Einsatz der Blackbox ohne Flankierung gestattet. Insoweit verbleibt es also bei Art. 22 DSGVO, denn die DSGVO bleibt gem. Art. 1 Abs. 5 lit. i DSA-KommE unberührt. Die in Art. 13 DSA-KommE vorgeschriebenen regelmäßigen Transparenzberichte müssen gem. Art. 23 Abs. 1 lit. c DSA-KommE allerdings immerhin „Indikatoren für die Genauigkeit der automatisierten Mittel" angeben und steigern so die Nachvollziehbarkeit. Sehr große Online-Plattformen[76] schließlich müssen nach Art. 29 Abs. 1 DSA-KommE beim Einsatz von Empfehlungssystemen (die wohl stets vorhanden sein dürften) deren „wichtigst[e] Parameter" dem einzelnen Nutzer offenlegen und ihm dabei wenigstens eine Option zur Verfügung stellen, die nicht auf Profiling im Sinne des Art. 4 Nr. 4 DSGVO beruht.

---

[75] Enquête-Kommission KI, Abschlussbericht, BT-Drs. 19/23700 (468).

[76] Art. 25 DSA-KommE legaldefiniert diesen Ausdruck anhand eines Grenzwerts von 45 Mio. Nutzern, den die Kommission durch delegierten Rechtsakt anhand der Bevölkerungsentwicklung der Union ändern können soll.

8. Umfassender lokaler Rechtsrahmen und Illusion seiner umfassenden Geltung

Transparenz- und Informationspflichten zur Verbesserung von Nachvollziehbarkeit und Überprüfbarkeit sind ein nicht unwichtiger Baustein der gegenwärtigen und kommenden KI-Regulierung.

III. Das Problem der grenzüberschreitenden KI

Zusammen mit den anderen Regelungen der geplanten EU-KI-VO und dann vielleicht noch gewürzt mit antidiskriminierungs-, gleichstellungs- und identitätspolitischen Zutaten klingt dies alles nach einem feuchten Traum für Regulierer und Normsetzer, nach der Erfüllung aller Wünsche. Doch wie beim Geschlechtlichen kommt es auf das Gegenüber an. Und das Gegenüber ist hier nicht so sehr das regulierte System oder der regulierte Anbieter, sondern andere Regulatoren jenseits der Grenze (s. u. 1.). Zwar gilt beim Zusammentreffen mehrerer Rechtsordnungen – anders als im Sexualstrafrecht – kein „Nein heißt nein". Doch muss man mit dem Gegenüber gleichwohl umgehen. Man kann entweder den Freuden und Kontakten ganz entsagen oder sich willenlos hingeben (s. u. 2.a)). Oder man kann seine Wünsche und Ziele einseitig durchsetzen (s. u. 2.b)). Man kann auch darauf warten, dass alle das gleich wollen (s. u. 2.c)). Oder man muss die wechselseitigen Bedürfnisse mühsam austarieren und Schranken verhandeln (s. u. 2.d)).

1. Ubiquität von KI und Räumlichkeit von Recht

Im Zeitalter der Digitalisierung und zukünftig auch der Künstlichen Intelligenz hat das Recht ein Problem: Staatliche Regelungen sind (notwendigerweise) lokal und jedenfalls an ein Staatsgebiet gebunden, während Algorithmen in der Cloud und durch die Vernetzung (potentiell) ubiquitär sind. Zwar ist das Internet kein rechtsfreier Raum, aber es ist auch nicht gerade ein Raum des Rechts, vor allem nicht der Raum (nur) *eines* Rechts. Auf Internet- und KI-Sachverhalte sind deshalb potentiell die Rechtsordnungen der etwa 200 Staaten[77] anwendbar.

Der Umstand, dass ein Sachverhalt Bezüge zu mehreren Rechtsordnungen hat, kann einen Juristen freilich nicht erschüttern. Er ist uns vertraut, seitdem die Menschen aus unterschiedlichen Rechtsordnungen miteinander in Kontakt treten, also seit Urzeiten (bzw. kurz danach).

[77] Wenn man die Staaten mit föderalem Rechtssystem – nicht nur etwa die USA, sondern hinsichtlich des Medienrechts auch Deutschland – hineinzählt, kommt man auf eine noch größere Zahl.

2. Regelungsstrategien

Hierfür gibt es vier Lösungsansätze: Entweder schottet man sich als Gesellschaft ab und vermeidet Kollisionslagen bereits auf der tatsächlichen Ebene. Oder man diktiert seine Regeln den anderen. Oder man entwickelt universelle Rechtsregeln. Oder man regelt das Verhältnis der unterschiedlichen Rechtsordnungen zueinander.

– Die Abschottung (a)) ist die simpelste Lösung. Auch folgt sie zwanglos aus dem Souveränitätsprinzip als der Grundregel des Völkerrechts und ist sozusagen die „Default Option" des Informationskollisionsrechts. Allerdings entspricht sie weder den Bedürfnissen einer mobilen Weltgesellschaft noch denen einer global vernetzten Wirtschaft.

– Ähnlich einfach ist die Welt für mächtige Staaten und Staatenverbünde. Sie tun einfach, was ihnen gefällt und zwingen ihren Willen, ihr Regelungsmodell und ihre Werte anderen Staaten, die sich nicht abschotten, auf (b)).

– Wenn man vom Völkerrecht her kommt, wäre ein universales Recht die schönste Lösung (c)). Es ist aber weder in Reichweite, zudem in Anbetracht der kulturellen Unterschiedlichkeit und Vielfalt auf unserem Planeten vielleicht auch nicht einmal wünschenswert.

– Der letzte Ansatz ist der der Koordinierung der Rechtsordnungen bzw. deren Anwendung. In der Rechtswissenschaft wird dies Kollisionsrecht genannt (d)). Hier wird es unübersichtlich und mühsam. Angesichts der praktischen Untauglichkeit der anderen Lösungen ist die Koordinierung der Rechtsordnungen aber alternativlos.

a) Abschottung und Öffnung

Die Souveränität der Staaten ist der Grundbaustein des Völkerrechts. Und Souveränität bedeutet (die Möglichkeit von) Grenzen und die freie Gestaltbarkeit der inneren Verhältnisse, der Gesellschafts- und Informationsordnung.

Deshalb dürfen – das ist der völkerrechtliche und internationalrechtliche Ausgangspunkt – Staaten ihr Verbraucherschutzsystem autonom und souverän gestalten. Dies schließt auch die Schließung von Grenzen ein, für Mensch und Maschinen.

Das kommende EU-KI-Regime, das ja in rechtssystematischer Hinsicht Teil des Produktsicherheitsrechts ist, kennt wie jedes Rechtsgebiet mit Außenhandelskomponente auch Modi der Einfuhrbeschränkung und (in der Konsequenz) von Zoll- und Grenzkontrollen; erkennbar ist dies an den ausdrücklichen Regelungen für Importeure (Legaldefinition in Art. 3 Nr. 6 EU-KI-VO-KommE) insbesondere von Hochrisikosystemen (Art. 26 EU-KI-VO-KommE). Speziell in dem Zusammenhang zu erwähnen ist das CE-Kennzeichen, das nicht nur für Geräte gilt (Art. 12 EU-MaschinenVO-KommE), sondern auch für KI-Systeme gelten soll (ausdr. für Hochrisikosysteme ErwGr. 67 EU-KI-VO-KommE).

Im Vergleich mit der DSGVO ist bemerkenswert, dass die EU-KI-VO-KommE den Import (von KI) regelt und beschränkt, während die DSGVO den Export (von personenbezogenen Daten) reglementiert.[78]

Und so wie die Schließung von Grenzen das souveräne Recht eines Staates ist, ist es auch deren Nicht-Schließung, also ihre Öffnung. Gerade aus einer kulturellen Bereicherungs- und Fortschrittsperspektive ist das natürlich die aufgeklärte und liberale, also eigentlich europäische Perspektive. Nun aber ist der europäische Fortschrittsglaube in Bezug auf Digitaltechnologien etwas gebremst, jedenfalls nicht unkritisch gegenüber dem disruptiven unternehmensgetriebenen Digitalkapitalismus des Silicon Valley oder dem autoritären Staatsdirigismus chinesischer Prägung. Eine Öffnung hat die Chance und den Preis, dass man sich ausländischem Einfluss aussetzt. Je nach gesellschaftlicher Disposition empfindet man das Fremde als Bereicherung oder Bedrohung.

Wie groß die Akzeptanz der Technik- und Freiheitsbeschränkungen bei uns Europäern ist, kann wohl nicht eindeutig gesagt werden. Einerseits ist das Datenschutzbewusstsein in Europa sicherlich höher als in anderen Teilen der Welt. Andererseits ist das Maß der unkritischen Nutzung von Whatsapp und Google-Diensten in der EU nicht viel niedriger als in anderen Teilen der Welt. Die Normgeber in Europa müssen sich also durchaus fragen (lassen), ob der verhältnismäßig restriktive Regulierungskurs in Technologiefragen, vor allem im Datenschutz, womöglich aber auch im Hinblick auf KI, den Mehrheitswillen der europäischen Bevölkerung trifft.

b) Aufzwingen der eigenen Regeln

Daneben sehen wir den Einsatz politischer und wirtschaftlicher Macht, um die eigenen Rechtsvorstellungen auf andere Rechtsordnungen zu übertragen und so divergierende Regelungen auszuschalten bzw. ihre eigene Rechtsordnung zur Anwendung zu bringen. Ein wirtschaftlicher Hebel kann das tatsächliche Innehaben von oder die Kontrolle über digitale Infrastrukturen sein, etwa über Soziale Netzwerke, Clouds und auch technische Standards (Internet).

Einen wirtschaftspolitischen Hebel nutzt die Europäische Union auf dem Feld des Datenschutzrechts mit dem Instrument der Angemessenheitsentscheidung, die vor allem wirtschaftlich schwächere Staaten dazu bringt, ihr Datenschutzrecht dem europäischen Vorbild anzugleichen („Brussels Effect"). – Auch die chinesische „Neue Seidenstraße" mag in diese Richtung wirken („Beijing Effect").[79]

Ausdrücklich sieht die Europäische Kommission in ihrem Regelungsvorschlag für eine KI-VO eine deutliche Stärkung („erheblich gestärkt") der Rolle der EU „bei der Formulierung weltweiter Normen und Standards sowie der Förderung ver-

---

[78] *Bomhard/Merkle*, RDi 2021, 276 (278 f. [Rn. 16]).

[79] Vgl. *Passi*, in: Xing (Hrsg.), Mapping China's ‚One Belt One Road' Initiative (2019), 167 (185 f.), der beschreibt, wie China die im Zusammenhang mit der Neuen Seidenstraße gemachten Schulden gezielt zur Einflussnahme ausnutzt.

trauenswürdiger KI, die mit den Werten und Interessen der Union in Einklang stehen“[80]. Auch sollen die neuen Regelungen „in nichtdiskriminierender Weise für Anbieter von KI-Systemen [gelten,] unabhängig davon, ob sie in der Union oder in einem Drittland niedergelassen sind“ (ErwGr. 10 EU-KI-VO-KommE); hier wird einseitige Normsetzung[81] mit dem schönen Wort der Diskriminierungsfreiheit verbrämt.

c) Weltweite Harmonisierung

Mit der Herausbildung eines „Rechts der Künstlichen Intelligenz“ wird rechtspolitisches Mittel- und Fernziel die Entwicklung eines universalen KI-Rechts sein.[82] Hierfür fehlt es aber bislang noch an Herausbildung von nationalen KI-Rechtsordnungen, geschweige denn deren Vergleichbarkeit.

Mit den USA wird eine „Agenda für den globalen Wandel“ angestrebt, was ein transatlantisches KI-Abkommen einschließen würde und einen multilateralen Ansatz bei der KI-Regulierung anstrebt.[83] Gegenüber anderen Rechtsordnungen, insbesondere der China, findet aber in diesem Schritt gerade keine Koordinierung statt, sondern wirtschafts- und digitalpolitisch gerade eine Abgrenzung.

d) Kollisionsrecht

Das heutige moderne Kollisionsrecht hat sich über die letzten bald 200 Jahre entwickelt. Im Zusammenspiel der Rechtsordnungen der Welt haben sich für viele Sachverhaltskonstellationen Regeln herausgebildet, die die Staaten und Rechtsordnungen wechselseitig und reziprok akzeptieren. Kollisionsrecht ist freilich nicht universelles Recht, sondern hat sich in Koordination der Rechtsordnungen entwickelt (und entwickelt sich auf diesem Wege auch noch weiter). Auch ist der Souveränitätsgedanke dem Kollisionsrecht nicht fremd, sondern in Gestalt des Ordre public-Vorbehalts durchgehend präsent.

---

[80] EU-Kommission, 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 (COM(2021) 206 final, Ziff. 1.3. [S. 6]).

[81] Abwägend-kritisch zu diesem Ansatz der EU *Heldt/Hennemann*, Die Goldenen Zwanziger Jahre der Technikregulierung, bidt-Blogbeitrag v. 3.8.2021, https://www.bidt.digital/blog-technikregulierung/.

[82] Übersicht bei *Ebers*, in: Ebers/Heinze/Krügel/Steinrötter, Künstliche Intelligenz und Robotik, 2020, § 3 I.

[83] Gemeinsame Mitteilung von *Europäischer Kommission* und *Hohem Vertreter der Union für Außen- und Sicherheitspolitik*, „A new EU-US agenda for global change“ v. 2.12.2020, JOIN(2020) 22 final.

aa) Anknüpfungen

Das Kollisionsrecht beruht auf dem Gedanken der nächsten oder passendsten Anknüpfung (Genuine Link). Die Rechtsordnung mit der größten Nähe zum Sachverhalt soll diesen entscheiden können. Hierfür und für unterschiedliche Konstellationen gibt es unterschiedliche Anknüpfungen: Rechtswahl, Territorialität, Marktort, Staatsangehörigkeit …

Wegen seiner Rückbindung an ein staatliches und damit örtliches Recht ist Kollisionsrecht immer im Raum verortet. Das passt auf die digitale Ubiquität (s. o. 1.) grundsätzlich nicht. Das Recht kann hier (bislang) nur zu Behelfsanknüpfungen greifen. So wird auf den Sitz der betroffenen Parteien abgestellt oder einen Marktort oder man verpflichtet zur Speicherung der Daten an einem bestimmten Ort (Data Localisation). Denkbar wäre auch, auf die Staatsangehörigkeit (personales Schutzprinzip, „Bürgerschutz") oder den „Ursprung" der Daten abzustellen.

In Anbetracht der digitalen Natur von KI-Anwendungen ist der sachlich-räumliche Anwendungsbereich[84] der EU-KI-Regulierung weit[85] (ErwGr. 11 EU-KI-VO-KommE). Die Regelungen zum räumlichen Anwendungsbereich lehnen sich an das Vorbild des Art. 3 DSGVO an und entsprechen damit dem gegenwärtig gängigen Regulierungsansatz auf europäischer Ebene. So sollen die KI-Regelungen dem Marktortprinzip (Art. 2 Abs. 1 lit. a KI-VO-E: „Anbieter, die KI-Systeme in der Union in Verkehr bringen oder in Betrieb nehmen"), dem personalen Schutzprinzip (Art. 2 Abs. 1 lit. b EU-KI-VO-KommE[86]: „Nutzer von KI-Systemen, die sich in der Union befinden") und dem Erfolgsortprinzip (Art. 2 Abs. 1 lit. c EU-KI-VO-KommE: „das vom System hervorgebrachte Ergebnis in der Union verwendet wird") folgen.

bb) Einseitige Rechtsanwendungsregel oder echtes Kollisionsrecht?

Kollisionsrecht funktioniert aber nur dann, wenn Realphänomene vergleichbar kategorisiert werden, wie etwa das familiäre Zusammenleben von Menschen (Ehe), die Verfügungsmöglichkeit über Güter (Eigentum, Besitz) oder Verletzungshandlungen (Delikt). Bei Künstlicher Intelligenz aber fehlt es an einer solchen Kanonisierung, denn sie unterscheiden sich grundlegend schon darin, ob es sich diesbezüglich um eine Willenserklärung, Datenverarbeitung oder künftig gar eine elektronische Person oder jedenfalls zurechenbares Vertreterhandeln handelt. Auch wird nicht jede Rechtsordnung und auch nicht jede Epoche der Kategori-

---

[84] Anders als etwa noch die DSGVO (in deren Art. 2 und 3) unterscheidet Art. 2 EU-KI-VO-KommE nicht mehr zwischen sachlichem und räumlichem Anwendungsbereich.

[85] *Bomhard/Merkle*, RDi 2021, 276 (278 [Rn. 12]).

[86] Wobei sich „location" wegen der unkörperlichen Natur von Software (und damit auch KI) nur auf die Hardware bezieht, auf denen solche Software läuft (*Bomhard/Merkle*, RDi 2021, 276 [278, Rn. 13]). Wie der Rechtsakt verteilte und vernetzte Systeme (KI as a Cloud) insoweit behandelt wissen will, ist unklar (a. a. O., Rn. 14).

sierung des KI-Rechts als in systematischer Hinsicht Produktsicherheitsrecht[87] folgen wollen.

Die EU-KI-Regulierung möchte risikobasiert sein (vgl. ErwGr. 14 EU-KI-VO-KommE). Doch wie soll ein Kollisionsregime für Risikobasierung aussehen, wenn die Risiken in unterschiedlichen Rechtsordnungen und Kulturkreisen unterschiedlich gewichtet und beurteilt werden? So geht es nicht nur um die Eingruppierung entsprechend der Kritikalität, sondern um das Konzept von Kritikalität überhaupt. Die Klassifizierung als Hochrisikosystem in der EU stellt auf Sozialleistungen (ErwGr. 37 EU-KI-VO-KommE), Sicherheitsbehörden (ErwGr. 38 EU-KI-VO-KommE), Justiz (ErwGr. 40 EU-KI-VO-KommE) und Migration (ErwGr. 39 EU-KI-VO-KommE) ab, was mit Wertungen anderer Rechtsordnungen und Kulturräume nicht unbedingt deckungsgleich sein muss.

Immerhin international anschlussfähig hinsichtlich der Kategorie ist die Transparenz von KI und dessen Ergebnissen, weil sie kognitive Fähigkeiten des Menschen adressiert, die universell sind. Insoweit könnte man durchaus ein Kollisionsregime hinsichtlich der Frage sich vorstellen, ob Nachvollziehbarkeit oder bloße Erklärbarkeit genügt. Ebenfalls kollisionsrechtlich miteinander in Bezug gesetzt werden könnten Verfahrenspflichten (Risiko- und Qualitätsmanagement, Registrierung, Datennutzung, [Nutzer-]Transparenz, Dokumentation, Kontrolle durch Menschen [„Human in the Loop"], IT-Sicherheit, Fehlerfreiheit, Überwachungssysteme, Meldepflichten), daneben wäre auch die Güte der Trainingsdaten als eine mögliche technikrechtliche Pflicht zu bedenken.

cc) Defizite

Wegen der unzureichenden Passung kollisionsrechtlicher Ansätze für grenzüberschreitende Digitalsachverhalte nehmen die Staaten dann häufig unilateral auf diese Sachverhalte Zugriff. Dies geschieht über administrative Regelungen, für die dann nach den Prinzipien des Internationalen Verwaltungsrechts die Territorialität maßgeblich ist. Oder sie berufen sich auf den Ordre public, indem „Digitale Souveränität" zu einem überragend wichtigen Rechtsgut erklärt wird.

Konzeptionell wird das herkömmliche Kollisionsrecht aber wegen der Ubiquität digitaler Daten (s. o. 1.) überdehnt. Es ist darauf berechnet, einen ausnahmsweisen Bezug zu mehr als einer Rechtsordnung einer Lösung zuzuführen. Nicht gedacht aber ist es für den Fall, dass auf *alle* Fälle (potentiell) *alle* Rechtsordnungen anwendbar sind. In Bezug auf Digitalsachverhalte sind aber zudem nicht nur diese allgemeinen Randbedingungen des Kollisionsrechts problematisch, sondern es fehlt auch an einem allgemeinen weltweiten Verständnis über ein Daten- und Informationskollisionsrecht überhaupt.

---

[87] *Schallbruch*, DuD 2021, 438 (443).

3. Skizze eines KI-Kollisionsrechts

Es stellt sich also die Frage, wie ein „echtes" KI-Kollisionsrecht aussehen und funktionieren kann.

a) Kleinteiliges Kollisionsrecht als Anfang

Kollisionsrechtler müssen wir uns als glückliche Menschen vorstellen, wie sie sisyphosgleich den Stein der immer besseren Interoperabilität der Rechtsordnungen dieser Welt herumrollen. Auf dieser Mikroebene kann das hier zu verhandelnde Problem der Nachvollziehbarkeit und Überprüfbarkeit von KI-Systemen durchaus verortet werden.

Wie gut solche Systeme nachvollziehbar und überprüfbar sind, scheint wegen der gleichmäßigen Verteilung menschlicher Intelligenz und kognitiver Fähigkeiten (s. o. 2.d)aa)) über die Erde möglich; dass hier Bildungsunterschiede zu berücksichtigen sein werden, liegt auf der Hand.

b) Gesamthaftes Kollisionsrecht

Die fehlende Vergleichbarkeit und damit die Unmöglichkeit einer einfachen kollisionsrechtlichen Anknüpfung kann aber durch einen KI-Regimevergleich (Begriff in Anlehnung an *Teubner* und *Fischer-Lescano*[88]) ersetzt werden und ist für das Regulierungsrecht etwa von *Hannah Buxbaum* vor kurzem durchexerziert werden.[89] Hierfür werden Rechtsordnungen gesamthaft und funktional miteinander verglichen. Es werden nicht nur einzelne materielle Rechtsnormen in den Blick genommen, sondern auch Technikrecht, Verfahrensregeln und Durchsetzungsmöglichkeiten. Eine praktische Ausprägung für eine solche gesamthafte Betrachtung ist das Vorgehen bei den datenschutzrechtlichen Angemessenheitsentscheidungen der Europäischen Kommission. Methodisch wird dabei die funktionale Rechtsvergleichung auf das Kollisionsrecht übertragen.

Man kann insoweit von einer „Interoperabilität" der jeweiligen Regelungsregime sprechen. Dem entspricht auch, dass in der EU auch die KI, ähnlich wie Datenschutz und IT-Sicherheit, sektorübergreifend (und also gesamthaft) reguliert werden soll.[90] Man kann also von einem gesamthaften Vergleich „guter digitaler Governance"[91] sprechen.

---

[88] *Fischer-Lescano/Teubner*, Regime-Kollisionen, 2006.

[89] *Buxbaum*, Public Regulation and Private Enforcement in a Global Economy: Strategy for Managing Conflicts (Haager Akademie für internationales Recht), 2017.

[90] So ausdrücklich *Schallbruch*, DuD 2021, 438 et pass.

[91] Zu diesem Konzept im Zusammenhang mit KI *Kastrop/Ponattu*, DuD 2021, 434 ff.

IV. Fazit

Wenn nicht nur Brüssel, sondern die ganze EU ein Raumschiff wäre, dann würden wir mit unseren Regelungen gut reisen – und mit den kommenden noch besser. Weil wir aber nicht alleine auf dem Planeten sind, muss wegen der Ubiquität digitaler Sachverhalte die internationale Dimension mitgedacht werden. Dies tut die EU nicht, vielleicht weil sie sich wirtschaftspolitisch stark genug und moralisch überlegen fühlt. Dieser blinde Fleck wird uns dann aber künftig noch arg zu schaffen machen. Denn wenn wir nicht stark genug sind und uns nur darauf verlassen, auf der richtigen Seite der Geschichte des Informationsrechts zu stehen, kann es sein, dass wir als der „globale Osten" des digitalen Zeitalters enden.

Autorenverzeichnis

Prof. Dr. *Ziawasch Abedjan*, Institut für Praktische Informatik, Fachgebiet Datenbanken und Informationssysteme an der Leibniz Universität Hannover und Mitglied des L3S Research Center

Prof. Dr. *Christian Armbrüster*, Professur für Bürgerliches Recht, Handels- und Gesellschaftsrecht, Privatversicherungsrecht und Internationales Privatrecht an der Freien Universität Berlin

Prof. Dr. *Bettina Berendt*, Institut für Telekommunikationssysteme, Fachgebiet Internet und Gesellschaft an der Technischen Universität Berlin und Direktorin am Weizenbaum-Institut für die vernetzte GEsellschaft

Prof. Dr. *Philipp Hacker*, LL.M. (Yale), Professur für Recht und. Ethik digitaler Gesellschaft an der European New School of Digital Studies der Europa-Universität Viadrina Frankfurt/Oder und Research Fellow am Weizenbaum-Institut für die vernetzte Gesellschaft

Prof. Dr. Dr. *Eric Hilgendorf*, Lehrstuhl für Strafrecht, Strafprozessrecht, Rechtstheorie, Informationsrecht und Rechtsinformatik an der Julius-Maximilians-Universität Würzburg

Prof. Dr. *Gerrit Hornung*, LL.M., Professur für Öffentliches Recht, IT-Recht und Umweltrecht an der Universität Kassel und Direktor am Wissenschaftlichen Zentrum für Informationstechnik-Gestaltung (ITeG)

Lukas Hundertmark, Doktorand und wissenschaftlicher Mitarbeiter am Lehrstuhl für Bürgerliches Recht, Zivilprozessrecht und Handelsrecht an der Martin-Luther-Universität Halle-Wittenberg

Prof. Dr. *Ruth Janal*, LL.M., Lehrstuhl für Bürgerliches Recht, Immaterialgüter- und Wirtschaftsrecht an der Universität Bayreuth

Prof. Dr. *Rüdiger Krause*, Lehrstuhl für Bürgerliches Recht und Arbeitsrecht an der Georg-August-Universität Göttingen

Prof. Dr. *Anne Lauber-Rönsberg*, LL.M., Professur für Bürgerliches Recht, Immaterialgüterrecht, insb. Urheberrecht, sowie Medien- und Datenschutzrecht an der Technischen Universität Dresden

Prof. Dr. *Kai von Lewinski*, Lehrstuhl für Öffentliches Recht, Medien- und Informationsrecht an der Universität Passau

Prof. Dr. *Caroline Meller-Hannich*, Lehrstuhl für Bürgerliches Recht, Zivilprozessrecht und Handelsrecht an der Martin-Luther-Universität Halle-Wittenberg

Jan-Laurin Müller, Wissenschaftlicher Mitarbeiter am Lehrstuhl für Bürgerliches Recht, Immaterialgüter- und Wirtschaftsrecht an der Universität Bayreuth

Felix Neutatz, Wissenschaftlicher Mitarbeiter am Institut für Softwaretechnik und Theoretische Informatik an der Technischen Universität Berlin

Prof. Dr. Dr. *Frauke Rostalski*, Lehrstuhl für Strafrecht, Strafprozessrecht, Rechtsphilosophie und Rechtsvergleichung an der Universität zu Köln

Prof. Dr. *Giesela Rühl*, LL.M. (Berkeley), Lehrstuhl für Bürgerliches Recht, Zivilverfahrensrecht, Europäisches und Internationales Privat- und Verfahrensrecht und Rechtsvergleichung an der Humboldt-Universität zu Berlin

Prof. Dr. *Ute Schmid*, Professur für Angewandte Informatik insb. Kognitive Systeme an der Otto-Friedrich-Universität Bamberg und Gruppenleiterin der Fraunhofer IIS Projektgruppe Comprehensible AI

Prof. Dr. *Lauri Wessel*, Professur für Information Management und Digitale Transformation an der European New School of Digital Studies der Europa-Universität Viadrina Frankfurt/Oder