AI Regulation in Europe

Author: Dr. Philipp Hacker, LL.M. (Yale)*

Abstract:

With the regulation of Artificial Intelligence (AI), the European Commission is addressing one of the central issues of our time. However, a number of core legal questions are still unresolved. Against this background, the article in a first step lays regulatory foundations by examining the possible scope of a future AI regulation, and by discussing legal strategies for implementing a risk-based approach. In this respect, I suggest an adaptation of the Lamfalussy procedure, known from capital markets law, which would combine horizontal and vertical elements of regulation at several levels. This should include, at Level 1, principles for AI development and application, as well as sector-specific regulation, safe harbors and guidelines at Levels 2-4. In this way, legal flexibility for covering novel technological developments can be effectively combined with a sufficient amount of legal certainty for companies and AI developers. In a second step, the article implements this framework by addressing key specific issues of AI regulation at the EU level, such as: documentation and access requirements; a regulatory framework for training data; a revision of product liability and safety law; strengthened enforcement; and a right to a data-free option.

Contents:

* AXA Postdoctoral Fellow, Humboldt University of Berlin.

A. Introduction

When the decisive breakthrough for the training of deep neuronal networks was achieved in 2006,[1] it was hardly foreseeable that Artificial Intelligence (AI) techniques would find their way into the most diverse areas of economic and private activity within a few years: from credit scoring to personnel recruitment, from autonomous driving to medical diagnostics and research.

The European Commission has taken due heed of these developments. Alongside climate change, AI is set to become the second major topic of Ursula von der Leyen's presidency. With the White Paper on Artificial Intelligence, the Commission has now presented a much-anticipated blueprint for the promotion, but also for the regulation of this technology.[2] The White Paper presents a welcome opportunity to examine the possibility and the necessity of AI regulation at the European level. This contribution therefore briefly summarises, in all due brevity, the White Paper (B.) and lays regulatory foundations (C.) before looking at specific regulatory areas of AI regulation in Europe (D.), such as documentation and access requirements; a regulatory framework for training data; a revision of product liability and safety law; strengthened enforcement; and a right to a data-free option.

B. The structure of the Commission's White Paper on AI

On 19 February 2020, the European Commission unveiled three core documents for its digital agenda: a data strategy,[3] a report on security and liability issues in emerging technologies,[4] and the White Paper on AI. The latter is characterized by a clear dichotomy of promotion and regulation. The novel labels for these different perspectives are those of the 'ecosystem of excellence' (promotion of AI) and the 'ecosystem of trust' (regulation of AI).[5] Overall, the White Paper attempts a difficult, but generally worthwhile, balancing act of making Europe a centre of AI development and application, while at the same time adequately addressing the risks of this technology and ensuring that European fundamental rights and values are adequately enforced.

However, the two major parts of the White Paper remain almost unconnected. The Commission starts with a clear and factually accurate description of the importance of AI for research, the economy and the society at large, stresses the need for the intelligent use of diverse data sources, and then sets out, on four pages, dimensions of enabling and promoting AI development, especially for SMEs. However, IP law is conspicuously absent from the entire discussion, and the Commission hence fails to mention that the new copyright exception for text and data mining in Art. 3 of the CDSM Directive,[6] which specifically excludes commercial research, may prove a significant burden for enabling AI development, especially in SMEs.[7] This represents a patent incoherence in the 'ecosystem of excellence'.

---

[1] Groundbreaking Hinton, Osindero and The, 'A fast learning algorithm for deep belief nets', 18 Neural Computation (2006), 1527; overview in Goodfellow, Bengio and Courville, *Deep Learning*, 2016, 18.

[2] European Commission, On Artificial Intelligence - A European approach to excellence and trust, White Paper, COM(2020) 65 final.

[3] European Commission, A European strategy for data, Communication, COM(2020) 66 final.

[4] European Commission, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64 final, 8 et seq.

[5] European Commission (n 2), 3.

[6] Directive (EU) 2019/790.

[7] Ducato and Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment', 50 *IIC* (2019), 649, 666; Geiger, Frosio and Bulayenko, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Briefing for the JURI committee of the European Parliament, 2018, 20 et seq.; Margoni and Kretschmer, 'The Text and Data Mining exception in the Proposal for

The following 17 pages of the White Paper then offer a panorama of regulatory options, the details of which will be analysed below. Overall, the Commission is committed to a risk-based regulatory approach,[8] like the GDPR[9] and the High-Level Expert Group on AI.[10] This includes the acknowledgement of risks to fundamental rights such as privacy, data protection and non-discrimination,[11] but also, quite rightly, of risks to data quality, IT safety and security, as well as difficulties in law enforcement.[12] However, this second, regulatory part of the White Paper barely explicitly mentions the risks for innovation that regulation may entail, be it through direct regulation or indirectly via the existence of intellectual property rights or the data protection framework. Innovation issues are almost exclusively addressed in the first, the 'enabling' part of the White Paper. However, if Europe wants to be at the forefront of AI research and development, risks for innovation must also be considered as a crucial factor when adapting the regulatory burden. This will be spelt out in greater detail in the following sections.

C. Regulatory foundations

Before specific regulatory challenges raised by AI can be addressed, however, basic regulatory questions must be answered, which are also addressed in the White Paper at the beginning of its regulatory section.

I. Review of the legal portfolio and market solutions

The Commission quite rightly emphasizes at the outset that existing EU and Member State law already contains a number of regulatory instruments which are – essentially – formulated in a technology-neutral way (e.g. the GDPR, anti-discrimination, unfair commercial practices, product liability law).[13] Therefore, any regulatory strategy must begin with a detailed examination of the extent to which the relevant risks are already covered by existing law. This can only be sketched in this contribution.[14] Overall, however, the Commission rightly notes that in some areas, there is at least a need for selective adjustments (see below, D.III. and D.V., concerning product liability/safety and data protection law). Furthermore, a broader case for reform can be made particularly in the areas of documentation requirements, of a regulatory framework for training data and of law enforcement (see below, D.I., II. and IV.).

It is also important to recognise that there is no need for regulation where adequate market solutions exist to address the relevant risks.[15] This, too, needs to be examined on a sector- and

---

a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs', Working Paper, 2018, 4 et seq.

[8] European Commission (n 2), 17 et seq.

[9] See, e.g., Lynskey, *The Foundations of EU Data Protection Law*, 2015, 81 et seqq.; *Gellert*, 'Data protection: a risk regulation?', 5 International Data Privacy Law 2015, 3; Article 29 Data Protection Working Party, 'Statement on the role of a risk-based approach in data protection legal frameworks', WP 218, 2014, 2.

[10] High-Level Expert Group on Artificial Intelligence, Policy and Investment Recommendations for Trustworthy AI, 2019, 37 et seq.; in a similar vein also German Data Ethics Commission, Opinion of the Data Ethics Commission, 2019, 173 et seqq.

[11] European Commission (n 2), 11 et seq.

[12] European Commission (n 2), 12 et seq.

[13] European Commission (n 2), 13 et seq.

[14] For a more comprehensive analysis in this respect, see, e.g., Mazzini, A System of Governance for Artificial Intelligence through the Lens of Emerging Intersections between AI and EU Law, Working Paper, 2019, https://ssrn.com/abstract=3369266; see also Hacker, 'A Legal Framework for AI Training Data', Working Paper, 2020.

[15] Veljanovski, 'Economic Approaches to Regulation', in: Baldwin, Cave and Lodge (Eds.), *The Oxford Handbook of Regulation*, 2010, 18, 21 et seq.

application-specific basis. Particular attention should be paid to the extent to which there are sufficient incentives for developers to consider certain risks at the level of the algorithmic model itself.

II. Scope of application of an 'AI Regulation'

Beyond extant EU law, any technology-specific 'AI regulation' will initially be faced with the crucial challenge of defining its own area of application in a way that is open to technical development and at the same time operationalisable for regulatory addressees.[16] Here lies perhaps the greatest weakness of the White Paper. The Commission, in discussing this point, merely provides a number of examples of AI applications and processes and, ultimately, refers to the definition of the High-Level Expert Group, which itself is rather vague, unclear and ultimately unfit for legal purposes.[17] However, these semantic difficulties point to a genuine problem: basically every technical book on AI uses its own definition so that even computer scientists have not yet been able to agree on a uniform concept of 'AI'.[18]

In my view, there are two alternative solutions to this problem. First, it seems advisable to speak, instead of AI, of machine learning (ML) techniques, which are much more clearly defined. Mitchell's definition is essentially agreed upon in the technical literature:[19] 'A computer program is said to learn from experience *E* with respect to some tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.'[20]

This terminological change from AI to ML not only makes the scope of application more tractable; it also reduces the pathos sometimes associated with the evocation of 'Artificial Intelligence', and mitigates inflated expectations as well as unrealistic risk attributions evoked by the term 'AI'. However, such a shift implies that certain techniques, which form part of AI but not of ML, are excluded (e.g., *knowledge bases*).[21] This seems a step worth taking, nevertheless: the specific risks of AI (unpredictability, manipulability, technical autonomy, lack of transparency, lack of training data quality, discrimination),[22] also mentioned by the Commission,[23] typically do not occur outside ML, or in any case only to an extent comparable to classical software. What the EU needs, therefore, is an 'ML Regulation', not an 'AI Regulation'. The former can be tailored more precisely to the regulatory risks.

A second, alternative possibility would be to reduce the degree of technological specificity by introducing regulation not for AI/ML, but for *software* associated with the above-mentioned risks typical for ML. These risks would then have to be defined in more detail. However, compared to a determination of the scope via the concept of ML, the second option faces the disadvantage that recourse to specific risks entails a considerably greater degree of legal uncertainty. It seems preferable, therefore, to delineate the scope by reference to ML, and to

[16] Scherer, 'Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies', 29 Harv. JL & Tech. (2015): 353, 396.
[17] High-Level Expert Group on Artificial Intelligence, A Definition of AI, 2019, 6.
[18] Overview for example in Russell and Norvig, *Artificial Intelligence*, 3rd ed. 2010, 1-5; see also Parnas, 'The real risks of artificial intelligence', 60(10) Communications of the ACM (2017), 27, 27.
[19] See only Goodfellow, Bengio and Courville (n 1), 96 et seqq.
[20] Mitchell, *Machine Learning*, 1997, 2.
[21] Goodfellow, Bengio and Courville (n 1), 2, 9.
[22] In greater detail Scherer (n 16), 362 et seqq.; Müller (ed.), *Risks of Artificial Intelligence*, 2016; Mazzini (n 14) 9 et seq.; Zech, 'Risiken digitaler Systeme', Weizenbaum Series #2, 2020, 24 et seqq., 44 et seqq., 50 et seqq.
[23] Cf. European Commission (n 2), 11-13.

take the degree of risk into account when designing the depth of regulation. In certain, justified cases, discussed in more detail below, the scope may even be extended to software in general (Product Liability Directive, see below, D.III.). In general, however, the novel regulatory framework should be restricted to ML.

III. Horizontal, vertical and diagonal regulation

Another basic question for the architecture of an ML regulation is whether it should be designed horizontally or vertically. The currently existing regulatory framework, which is not specific to AI, is largely of a horizontal nature (especially the GDPR, product liability and general product safety regulation, consumer protection and anti-discrimination law). This seems generally adequate for capturing the dynamic, rapidly evolving field of AI applications.

For new, ML-*specific* forms of regulation, however, a combination of horizontal and vertical elements appears advisable. In this endeavour, one area that has also been the subject of major regulatory efforts in the recent past may serve as a regulatory role model: capital markets law. There, the so-called Lamfalussy procedure was introduced, which has proved to be an effective instrument for fast and precise legislation.[24] It consists of regulation at four levels, ranging, in its original version, from general framework conditions (Level 1) to more specific provisions (Level 2), technical refinements in guidelines (Level 3) and enforcement acts (Level 4).[25] In this way, legislation can be dealt with in a complex, multi-level format, drawing on the technical competence of various stakeholders. At the same time – and this is also a lesson from capital markets law – it is important to avoid regulatory over-determination of the subject area,[26] and to ensure sufficient democratic legitimation of the regulators. The Lamfalussy procedure will therefore need to be adapted.

To the extent possible, the horizontal and vertical elements of the different levels of regulation should enter into force at the same time, or with little delay. In this respect, one could speak of 'diagonal regulation'. Thus, developers may obtain concrete compliance perspectives at an early stage, and are not stuck with mere principles for longer than necessary. In this way, the approach can arguably combine flexibility with sufficient legal certainty.

1. Level 1 regulation: Principles, training data, and enforcement

New, ML-specific forms of regulation should contain a horizontal level (Level 1), at which, first and foremost, general principles are laid down (e.g. transparency, non-discrimination, data quality, robustness, IT security).[27] This follows a tradition of principles-based regulation in areas subject to dynamic evolution, particularly, again, in capital markets law.[28] Those principles can draw on existing ones, such as those enshrined in Art. 5 GDPR and in anti-discrimination legislation. Including even these principles in a horizontal ML regulation instrument nevertheless makes sense, as this removes any doubt that these principles also apply

[24] See, e.g., Moloney, *EU Securities and Financial Markets Regulation*, 2015, 866 et seqq.
[25] Final Report of the Committee of Wise Men on the Regulation of European Securities Markets, 2001, 6 [henceforth Lamfalussy Report].
[26] Critical of the Lamfalussy procedure in this respect, for example Ferrarini, 'Contract Standards and the Markets in Financial Instruments Directive (MiFID)', 1 European Review of Contract Law 2005, 19, 31 et seq.; Ferran, *Building an EU Securities Market*, 2004, 84 et seq.
[27] See also Scherer (n 16), 394.
[28] See, e.g., Black, 'Forms and paradoxes of principles-based regulation', 3 Capital Markets Law Journal (2008), 425

to machine learning regardless of the individual conditions for the application of the GDPR (Art. 2(1), 4(1) GDPR: personal data)[29] and anti-discrimination law (significant restrictions of scope),[30] subject only to possible exceptions at Levels 2-4.

Furthermore, these principles should be combined with clauses pegging them to the technical state of the art. This would mean that all of the principles need to implemented *by design*, to the extent possible (cf. Art. 25 GDPR). For example, in order to prevent unjustified discrimination, measures of algorithmic fairness, i.e. the reduction of discrimination at the level of the ML model itself, must generally be applied in accordance with current best practices.

Furthermore, overarching rules on a number of issues addressed below (D.) ought to be located at a Level 1: rules on ML documentation, training data, product liability and safety as well as enforcement. Matters of enforcement were placed, in the original Lamfalussy process, at Level 4,[31] but this would fail to do justice to the importance, and difficulties, of enforcement in the field of machine learning.

Finally, certain aspects of ML applications that trigger particularly high risks, regardless of their sectoral use, could be partially regulated on a horizontal basis, with refinements at Levels 2-4. The example of remote biometric identification mentioned by the Commission White Paper is relevant here,[32] as this technology is of considerable importance for data protection, non-discrimination and access to economic goods and public infrastructure.

2. Level 2 regulation: Sector specificity

To render horizontal principles and rules more concrete, further regulation should be introduced on a sector-specific basis (Level 2), if, and only if, there is a real need for them in addition to existing sectoral legislation. Level 2 regulation can be more precisely tailored to different risks,[33] but also to market conditions. In particular, the possible existence of specific market solutions, which could speak against ML regulation, can be taken into account here. The risk-based approach will therefore be primarily applied at this second level. The sectors of medicine, autonomous driving, recruiting and personalised advertising seem to merit particularly urgent consideration, as (i) ML is already used in these areas and (ii) empirical studies have documented the poor quality and/or susceptibility to discrimination of some of these applications.[34] In addition, the provisions for specific, horizontally-regulated techniques (e.g. remote biometric identification), must be concretised in a sector- and application-specific manner at Level 2.

---

[29] Ostveen, 'Identifiability and the applicability of data protection to big data', 6 International Data Privacy Law (2016), 299, 307.

[30] See in detail Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law', 55 Common Market Law Review (2018), 1143, 1154 et seqq.; for US law, see Barocas and Selbst, 'Big Data's Disparate Impact', 104 California Law Review (2016), 671, 694 et seqq.

[31] Lamfalussy Report, 6, 40.

[32] European Commission (n 2), 18, 21 et seq.

[33] Cf. Mazzini (n 14) 14.

[34] Obermeyer et al., 'Dissecting racial bias in an algorithm used to manage the health of populations', 366 Science (2019), 447; Wilson, Hoffman and Morgenstern, 'Predictive inequity in object detection', Working Paper, 2019, https://arxiv.org/abs/1902.11097; Lowry and Macpherson, 'A blot on the profession', 296 British Medical Journal 1988, 657; Sweeney, 'Discrimination in Online Ad Delivery', 56(5) Communications of the ACM 2013, 44.

Finally, the state-of-the-art clauses can also be described in more detail at this level. To mitigate discrimination, for example, a large and continuously growing number of different methods and metrics exists.[35] As a general matter, given this dynamic research environment, it does not seem advisable to prescribe one of them specifically. However, normative specifications can be made for individual sectors or applications (e.g. rather individual fairness or rather group fairness;[36] rather an equalization, between groups, of the false positive rate or of the false negative rate[37]). In general, however, it must be sufficient if the developers use any legally effective and technically recognised method that is in line with the state of the art.

In contrast to the Lamfalussy procedure under capital markets law,[38] however, these regulations should in principle not only be adopted by means of delegated Commission regulations, but in the usual EU legislative procedure,[39] for reasons of democratic legitimacy and discursive openness. Institutional competence can be enhanced by means of expert testimony.

3. Level 3 regulation: Safe harbours

Principles-based and risk-based regulation affords the advantage of capturing dynamic processes. At the same time, however, it engenders considerable legal uncertainty for addressees,[40] which can lead to innovation risks, especially in the area of ML development. Therefore, application-specific safe harbors should be introduced, the observance of which would guarantee compliance with the respective Level 1 and 2 regulations. This mechanism is also well-known from capital markets law.[41] These safe harbors can be formulated by supervisory authorities, but better still by the legislator, possibly by means of delegated implementing regulations.

At this safe harbour level, specific metrics and measurable thresholds – developed in a process involving all stakeholders and reviewed on a regular basis – can be defined. For example, in the field of non-discrimination, these could be fairness metrics which implement a certain degree of so-called statistical parity. This measure describes how differently protected groups may be treated with regard to positive selection.[42] Similarly, certain performance metrics could be specified for data and predictive quality.[43] If developers wanted to use other metrics or values, they would be free to do so, but at their own risk.[44]

4. Level 4 regulation: Guidelines and self-regulation

The horizontal and vertical requirements can be further specified by (non-binding) guidelines. To this end, the establishment of specific, specialised authorities or committees would appear

[35] Overview at Dunkelau and Leuschel, 'Fairness-Aware Machine Learning', Working Paper, 2019.
[36] Zehlike, Hacker and Wiedemann, 'Matching Code and Law: Achieving algorithmic fairness with optimal transport', 34 Data Mining and Knowledge Discovery 2020, 163, especially 188 et seq.
[37] See, e.g., Kleinberg, Mullainathan and Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores', 67 ITCS (2017), Article 43.
[38] Lamfalussy Report, 6, 28.
[39] Different view (delegation of specific rule-making to AI agency) in Scherer (n 16), 395.
[40] Black (n 28), 426.
[41] Moloney (n 24), 699, 712, 727, 752, 857.
[42] Dwork et al., 'Fairness through Awareness', Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 2012, 214, 215, 218, also on limitations of this approach.
[43] See Deussen et al., Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model, DIN SPEC 92001-1, 2019.
[44] Cf. Black (n 28), 445.

to make sense. As is well known, the GDPR also provides for concretising mechanisms which draw on regulated self-regulation, under Article 40 GDPR (approved codes of conduct) and Article 42 GDPR (certification). Such procedures should also be part of any ML regulation.[45]

IV. Concretising the risk-based approach

A final general aspect concerns the risk-based approach itself. It affords the advantage to adapt the regulatory burden to the importance of the legal interests involved, to the expected damage and to the avoidability of risks by those affected.[46] In this way, regulatory costs for the addressees can be minimised in an innovation-friendly manner; simultaneously, proportionality can be safeguarded.

However, the concretisation of the risk-based approach poses a particular challenge. One possibility is the definition of risk classes. The German Data Ethics Commission, for example, proposes five of these,[47] the Commission seems to endorse only two (high-risk; other).[48] In any case, certain sectors or forms of application would have to be assigned to these classes. This is well-known from other areas of risk regulation,[49] such as medical devices.[50] However, it will be a tremendous challenge to formulate transparent, clearly defined allocation criteria across sectors. Furthermore, as the Commission rightly emphasises, any risk-adequate design of legal requirements must consider not only the sector concerned but also the specific use of the ML model.[51] This allows a certain degree of legal certainty (sector) to be combined with sufficient fine-tuning in individual cases (application). The risk classes therefore appear to make sense as heuristics, but may be too abstract for the vast field of ML regulation.

In my view, it will often be easier to *implicitly* adapt concrete regulations to the degree of risk than to *explicitly* assign sectors or applications to concrete risk classes. An example: In the field of ML recruiting, the user of an ML application could be obliged to provide candidates with the ten most important factors for their ML-based score or classification. In general, it should be easier to reach a legislative consensus on such concrete measures than on an abstract risk classification of 'ML recruiting', which would run the additional risk of being under-complex.

In many cases, at any rate, the risk specificity of ML regulation will have to be determined by the interpretation of *existing* law, including the general clauses and undefined legal terms contained in it. Guidelines by supervisory authorities, codes of conduct and certification mechanisms will, as seen, be indispensable to reduce legal uncertainty in this area, too.

D. Building blocks of ML regulation at the EU level

On the basis of these regulatory foundations, ML regulation will, in a first step, have to focus on a number of key issues. At the European level, the following five areas seem most urgent. They can form the building blocks of an integrated ML policy for Europe.

---

[45] Martini, *Blackbox Algorithmus*, 2019, 320 et seqq.
[46] German Data Ethics Commission (n 10), 173 et seq.
[47] German Data Ethics Commission (n 10), 177.
[48] European Commission (n 2), 17.
[49] See Adler, 'Risk, death and harm: The normative foundations of risk regulation' 87 Minn. L. Rev. (2003), 1293, 1348 et seqq.; Cao and Rieger, 'Risk classes for structured products' 9 Annals of Finance (2013), 167.
[50] See Annex VIII of the Medical Devices Regulation (EU) 2017/745; Mazzini (n 14) 11 et seq.
[51] European Commission (n 2), 17 et seq.

I. Documentation and access

The documentation requirements set out in the White Paper may be considered as that part of the proposals on which legislative consensus will most likely be achieved, since they do not impose any material restrictions on ML developers.[52] In this, they resemble disclosure rules, a staple of EU market regulation. Concerning ML, it seems necessary to develop a specific documentation and access regime for internal processes at Level 1, which must go far beyond the documentation requirements in Article 30 GDPR.[53]

1. Documentation: Content and scope

Both for effective supervision and for liability litigation, internal documentation of the creation of an ML model (the ML pipeline) is necessary.[54] In particular, the documentation of the following aspects seems desirable:

- Provenance of, and metadata, about the training data (e.g., descriptive statistics);
- Type of pre-processing;
- Reasons for selecting the relevant input variables (features);
- Reasons for selecting the type of machine learning model (linear regression, decision tree, deep neural network, etc.);
- Hyperparameters and their tuning;
- Performance metrics common in the industry (e.g. accuracy, precision, recall, F1 score);
- Changes in the model over time, especially since the first field application;
- Historically deployed versions of the ML model, in an ex post inspectable way (version archiving);
- Input and output values of the application data, if necessary in anonymous form;
- The weights of the model (for deep neural networks: weights and biases).

The depth of documentation can be adapted to the risk profile of the ML model and its deployment (Levels 2-4).[55] Information that can be stored by simple logging should, however, be kept available for all ML models, since even with low-risk models damage, and thus liability litigation, cannot be excluded.

The possibility of documenting the weights of the individual features is both particularly important and controversial. As is well known, the technical representability of these weights varies considerably depending on the type of machine learning. While weights in linear or logistic regression can be easily and globally (i.e. for the entire model) extrapolated from the regression equation, the global determination of weights, in a way that can be interpreted by humans, is considerably more difficult or even impossible in random forests or deep neuronal

[52] European Commission (n 2), 19 et seq.
[53] Martini (n 45), 260 et seqq.
[54] See, in greater detail, Selbst and Barocas, 'The Intuitive Appeal of Explainable Machines', 87 Fordham Law Review 2018, 1085, 1130 et seqq.
[55] German Data Ethics Commission (n 10), 190.

networks.[56] Although different strategies of algorithmic ex ante and ex post transparency (e.g. perturbation analyses) can be applied here, the factors relevant to the decision can typically only be approximated locally (i.e. for individual decisions).[57]

In principle, therefore, weights ought to be documented for auditing purposes. Medium- and high-risk applications should additionally be obliged to create a list of the ten most important global features, with their corresponding weights. The length of the list could also be adapted depending on the addressees. If, for the technical reasons just described, it is not possible at all or only with unreasonable effort to establish this global list, such feature lists should be provided for representative (local) cases, with appropriate local weight distributions. In addition, such a list of key local factors should be drawn up for each actual decision in the field, insofar as this is possible at a cost commensurate with the risk.

2. Access authorization: Qualified transparency and individual access rights

In a second, equally important step, access to this documentation must be regulated. For reasons of legitimate business interests, an approach of qualified transparency is recommended here, which, in principle, forces disclosure only to actors obliged to keep results secret (e.g. supervisory authorities, auditors).[58] However, information not endowed with any significant confidentiality character should be made publicly available. This includes meta-information on training data (descriptive statistics) and quality criteria such as performance metrics. The Commission also seems to have this rightly in mind.[59] Another key issue is access in the context of liability litigation; rules on the burden of proof can help here (see below, D.IV.1.).

For applications above a certain risk level, it could also be provided that the affected persons themselves have a right of access to the list of key decision factors. In my view, such a claim can already arise from Article 15(1)(h) GDPR ('meaningful information'),[60] which, however, is only of limited help in view of the restricted scope of the provision (only 'automated decision-making' is expressly mentioned). In any event, the data subject's interest in information must be weighed against the interest in preserving trade secrets and the functionality of the model (*gaming the algorithm*[61]). In this respect, a balance could be achieved, for example, by omitting or disclosing in a deliberately imprecise way the numerical weights, while at the same time fully naming the corresponding features.[62]

---

[56] See, in greater detail, Lipton, 'The Mythos of Model Interpretability', 61(10) Communications of the ACM 2018, 36, 40 et seqq.

[57] See, e.g., Ribeiro et al., 'Why Should I Trust You', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), 1135; but see also Lapuschkin et al., 'Unmasking Clever Hans predictors and assessing what machines really learn', 10 Nature Communications (2019), 1, 2 (for average heatmaps).

[58] See also BGH, Case VI ZR 156/13, NJW 2014, 1235 para. 17 et seq., 33; cf. also Citron and Pasquale, 'The Scored Society: Due Process for Automated Predictions', 89 Wash. L. Rev. (2014), 1, 24-27.

[59] European Commission (n 2), 19.

[60] In the same vein Selbst and Powles, 'Meaningful information and the right to explanation', 7 International Data Privacy Law (2017), 233, 235 et seqq.; Mazzini (n 14), 50 et seq.; Dix, in: Simitis, Hornung and Spiecker gen. Döhmann, Datenschutzrecht, 2019, DSGVO, Art. 13 para. 17; on the whole, although more skeptical about Art. 13-15 GDPR, Wachter, Mittelstadt and Floridi, 'Why a right to explanation of automated decision-making does not exist in the general data protection regulation', 7 International Data Privacy Law (2017), 76.

[61] Bambauer and Zarsky, 'The Algorithm Game', 94 Notre Dame L. Rev. 2018, 1.

[62] In a similar vein Bäcker, in: Kühling and Buchner (eds.), DS-GVO, 2nd ed. 2018, Art. 13, para. 54.

3. Need for adaptation of existing legislation

The question of access is therefore partly a matter of a proper interpretation of the GDPR. This concerns access by individuals (Art. 15(1)(h) GDPR) and by authorities (Art. 58(1)(b, e-f) GDPR). However, there is a real need for new provisions at the European level (i) on access rights *to non-personal* data; (ii) on access rights *by other* auditors than data protection authorities (e.g. technical inspection agencies); and (iii) on the *documentation requirements* themselves. Article 30 GDPR is clearly insufficient in this respect.

II. A regulatory framework for ML training data

As suggested by the Commission,[63] a second key issue, partly intersecting with documentation rules, is the development of a regulatory framework for ML training data.[64]

1. The relevance of training data and the insufficiency of the GDPR

Training data are, from a technical perspective, foundational for ML, particularly for supervised learning and for reinforcement learning strategies.[65] In supervised learning, the ML model is built by making its predictions match ever closer the known values of the target variable, contained in the training data.[66] In reinforcement learning, the model devises an optimal strategy by learning from specific feedback (reward signals) received by the training environment.[67] This environment will often be a simulation environment; again, the data used to build this environment, often synthetic data,[68] is crucial for the success of the learning operation.[69] Arguably, the two key challenges in this respect, also mentioned by the Commission White Paper,[70] are data quality and discrimination prevention. A third important dimension, not covered here, is IT security.

Data protection law does provide for certain requirements in the area of data quality. For example, Article 5(1)(d) GDPR specifies that personal data must be 'accurate and, where necessary, kept up to date'. These prerequisites, however, are not only very vague and remain undefined in the GDPR. Even more importantly, the applicability of data protection law to training data is highly doubtful because it is often anonymized.[71] Hence, training data sets in supervised learning may often fall between the cracks of the GDPR. In training environments for reinforcement learning, its applicability is even more clearly excluded if synthetic data is used.

---

[63] European Commission (n 2), 18 et seq.
[64] See, in more detail, Hacker (n 14).
[65] On the distinction between supervised, unsupervised and reinforcement learning, see Russell and Norvig (n 18), 694 et seq.; Shalev-Shwartz and Ben-David, *Understanding Machine Learning*, 2014, 4 et seq.; Jordan and Mitchell, 'Machine learning: Trends, perspectives and prospects', 349 *Science* (2015), 255, 257 et seq.; Goodfellow, Bengio and Courville (n 1), 102 et seq.; Sutton and Barto, *Reinforcement Learning*, 2nd ed., 2018, 2.
[66] LeCun, Bengio and Hinton, 'Deep Learning', 521 *Nature* (2015), 436, 436 et seq.; Goodfellow, Bengio and Courville (n 1), 79 et seqq., 102.
[67] Sutton and Barto (n 65), 6; Jordan and Mitchell (n 65), 258.
[68] See on this Gallas et al., 'Simulation-based reinforcement learning for autonomous driving', Proceedings of the 36th International Conference on Machine Learning (2019), 1.
[69] van Wesel and Goodloe, 'Challenges in the Verification of Reinforcement Learning Algorithms', NASA/TM-2017-219628, 15; Sutton and Barto (n 65), 2.
[70] European Commission (n 2), 18 et seq.
[71] Ostveen (n 29).

2. Toward a discrimination-sensitive quality regime for training data

A regulatory framework therefore ought to emancipate itself from data protection law and should, in a horizontal Level 1 instrument, include a quality assurance regime that is sensitive to discrimination and at the same time takes into account innovation risks. For this purpose, regulation may draw on an extensive literature in computer science dealing with data quality metrics[72] and discrimination prevention in data sets.[73] From this literature, five key aspects for a discrimination-sensitive quality regime for training data emerge that seem particularly relevant from a regulatory perspective.

(1) Accuracy: It is consented in the computer science literature that data used to build data-driven models should be accurate,[74] as mentioned by Article 5(1)(d) GDPR. However, there are different dimensions of accuracy (e.g., binary right-wrong or continuous distance from the correct value). Depending on the riskiness of the application, a smaller or larger margin of error might be permitted.[75]

(2) Timeliness: Another key feature is the timeliness of training data.[76] It is well-known that historical bias in training data may be perpetuated by ML models.[77] In this sense, yesterday's data must not determine tomorrow's choices. However, depending on the type of data, frequent updates may or may not be required.[78] Some measurements, for example the effectiveness of a certain medical drug, may remain quite constant over a longer period of time.

(3) Completeness and feature diversity: Furthermore, missing values are a common issue in data sets; completeness is, of course, the gold standard in this dimension.[79] However, it may also make a difference whether the features used for training the ML model, even if equipped with complete values, all measure more or less the same thing or are uncorrelated among one another (diverse).[80] To the extent that one feature closely correlates with membership in a protected group, adding other input variables less correlated with group membership may reduce the discriminatory impact of ML models. This is the reason behind rules such as the one found in German data protection law that scoring must not be based on address data only,[81] a feature known to correlate closely with ethnic origin in some areas.

---

[72] Overview by Lee et al., 'AIMQ: a methodology for information quality assessment', 40 Information & Management (2002), 133, 134 et seqq.; Heinrich and Klier, 'Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement', in: Hildebrand et al. (Ed.), Daten- und Informationsqualität, 4th ed., 2018, 47, 50 et seqq., each with an emphasis on completeness, accuracy, consistency and timeliness; fundamentally Wang, Storey and Firth, 'A Framework for Analysis of Data Quality Research', 7 IEEE Transaction on Knowledge and Data Engineering (1995), 623.

[73] See, e.g., Calders and Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures', in: Custers et al. (eds.), *Discrimination and Privacy in the Information Society*, 2013, 43; Romei and Ruggieri, 'A multidisciplinary survey on discrimination analysis', 29 The Knowledge Engineering Review (2014), 582.

[74] See only Heinrich and Klier (n 72), 55-57; Lee et al. (n 72), 134; Wang, Storey and Firth (n 72), 628 et seq.

[75] Cf. Information Commissioner's Office, Big data, artificial intelligence, machine learning and data protection, Version 2.2., 2017, para. 92.

[76] Heinrich and Klier (n 72), 59 et seq.; Lee et al. (n 72), 134; Hand and Henley, 'Statistical Classification Methods in Consumer Credit Scoring: a Review', 160 Journal of the Royal Statistical Society: Series A (Statistics in Society) (1997), 523, 525; Wang, Storey and Firth (n 72), 628 et seq.

[77] See Calders and Žliobaitė (n 82), 48 et seq.; Hacker (n 30), 1148.

[78] Information Commissioner's Office, 'Principle (d): Accuracy', https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/accuracy/.

[79] Heinrich and Klier (n 72), 52; Lee et al. (n 72), 134; Wang, Storey and Firth (n 72), 628.

[80] See Drosou and Pitoura, 'Multiple Radii DisC Diversity', 40 ACM Transactions on Database Systems (TODS) (2015), Article 4.

[81] § 31(1) no. 3 BDSG (German Data Protection Act).

(4) Group balance: Even if all values are accurate, up-to-date, complete and features diverse, certain protected groups may be underrepresented in the data set (sampling bias).[82] This may lead to a worse performance of the model with respect to that group, and to bias in the outcome of the ML model. Hence, data sets should be balanced between the different protected groups. This is rightly stressed by the Commission as well.[83]

(5) Representativeness: Finally, the training data or the training environment should be representative of the context.[84] This is crucial to allow the ML model to generalize from the test environment to real-world applications. One worst case, for example, would be a model used to steer an autonomous vehicle trained primarily on white persons; it may, in that case, fail or have difficulties to recognize people of a darker skin tone as human beings. A recently published study on object recognition suggests that precisely this is the case.[85] Such findings testify to the crucial importance of ensuring the representativeness of training data and training environments for the target context. The Commission is therefore right to highlight this aspect.[86]

3. The regulatory framework

Clearly, not every single ML model will have to fulfil all of the five mentioned quality parameters to the full extent; in fact, this will, in practice, often be impossible or entail prohibitive costs.[87] Therefore, following a risk-based approach, the burden of the quality regime must be adapted to the riskiness of the application (Levels 2-4). This means that, for example, the frequency of sample testing,[88] the intensity of error and bias correction algorithms,[89] and potentially even the establishment and duration of 'expiry dates' for certain data sets, must be targeted to the relevant context of application. If, for example, the right metrics for data accuracy are defined, one could link the tolerated margin of error of certain training data sets to the use the ML model is intended for (e.g., as a safe harbour). For high-risk applications, such as autonomous driving or medical diagnostics, developers would have to incur significantly greater costs to ensure training data quality than for a model helping to make appointments in a restaurant, for example.

III. Product liability and product safety law

Another key area for any type of ML regulation should be product liability and product safety.

---

[82] Calders and Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures', in: Custers et al. (eds.), *Discrimination and Privacy in the Information Society*, 2013, 43, 51; on sampling bias generally Hand, 'Classifier Technology and the Illusion of Progress', 21 Statistical Science (2006), 1, 8 et seq.

[83] European Commission (n 2), 19; European Commission (n 4), 8.

[84] Hand (n 82), 8 et seq.

[85] Wilson, Hoffman and Morgenstern (n 34).

[86] European Commission (n 2), 19; European Commission (n 4), 8.

[87] But see, for industry initiatives, Yang et al., 'Towards Fairer Datasets', Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*) (2020), 547; Google Research, 'Inclusive Images Challenge', kaggle (2018), https://www.kaggle.com/c/inclusive-images-challenge.

[88] Cf. Diakopoulos et al., 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms', Working Paper, 2018, Fairness, Accountability, and Transparency in Machine Learning, https://www.fatml.org/resources/principles-for-accountable-algorithms, under "Accuracy".

[89] See, e.g., Zehlike, Hacker and Wiedemann, 'Matching code and law: achieving algorithmic fairness with optimal transport', 34 Data Mining and Knowledge Discovery (2020), 163 and the overview in Dunkelau and Leuschel, 'Fairness-Aware Machine Learning', Working Paper, 2019.

1. Scope: Extension of the Product Liability Directive to software

The applicability of the Product Liability Directive 85/374/EEC (PLD) should be extended, as considered by the Commission,[90] to ML applications and (even stand-alone) software in general.[91] Due to their intangible character or their primary service elements, these may not fall under the concept of 'product' according to Article 2 PLD. In the absence of such a legislative extension, Article 2 must be applied by analogy.[92] Such an extension seems justified because the risks of, and the difficulties of proof arising from, the sub-standard production of ML applications, and of software in general, do not differ significantly from those arising from the production of tangible goods.

2. Defectiveness

Another core issue of product liability is to define criteria for the defectiveness of ML applications. Article 6 PLD links defectiveness to the legitimate safety expectations of users. How exactly these expectations can be defined is an ongoing matter of debate in product liability law. The most important test case will be design defects. Many scholars[93] and some Member State courts[94] hold that, for a design defect, it must generally be shown that an alternative product, which could have reasonably been developed by the manufacturer, would have been safer (risk-utility-test). For ML applications, however, not only are the general or specific safety obligations underspecified at the moment; at a conceptual level, it is also unclear what should be the exact object of the risk assessment.

For example, it may be the case that the entire fleet of a certain autonomous vehicle is, on average, safer than any other car type on the market (including those with human drivers), but that, nevertheless, it performs slightly worse than humans on some specific tasks, such as detecting stop signs. Suppose a single vehicle from this fleet misreads a stop sign and causes damage. This vehicle does not exhibit a manufacturing defect; rather, the probabilistic design of the image recognition ML unit leads to such accidents in some cases. Should this vehicle count as defective?

In this case, one may deny a design defect when looking at the average risk profile of the entire fleet because the ML unit, on average, functions better than any other device, including human eyes, and hence reduces accidents overall; or one may affirm a defect when looking at the individual vehicle's functioning concerning the specific task of stop sign detection. Overall, a fleet-based (systemic) perspective referring to overall risk seems preferable.[95] Generally

[90] European Commission (n 2), 15.
[91] Wagner, 'Robot Liability', Working Paper, 2018, https://ssrn.com/abstract=3198764, 11; Zech, 'Künstliche Intelligenz und Haftungsfragen', Zeitschrift für die gesamte Privatrechtswissenschaft (2019), 198, 212; but see also Schönberger, 'Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications', 27 International Journal of Law and Information Technology (2019), 171, 199; on product safety rules in this respect, see Wiebe, 'Produktsicherheitsrechtliche Pflicht zur Bereitstellung sicherheitsrelevanter Software-Updates', NJW 2019, 625, 626.
[92] Wagner (n 91), 11; Zech (n 91), 212.
[93] See, e.g., Howells, 'Defect in English Law', in: Fairgrieve (ed.), *Product Liability Law in Comparative Perspective*, 2005, 138, 143; Lenze, 'German Product Liability Law', in: Fairgrieve (ed.), *Product Liability Law in Comparative Perspective*, 2005, 100, 110 et seq.; Wagner (n 91), 12; but see the discussion in Whittaker, *Liability for Products*, 2005, 487 et seqq.
[94] See, e.g., BGH, Case VI ZR 107/08, NJW 2009, 2952, para. 18.
[95] Geistfeld, 'A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation', 105 California Law Review (2017), 1611, 1644-1647, 1651-1654; Wagner (n 91), 12 et seq.; Zech (n 91), 213.

speaking, at least during a transitional phase of AI adoption, an ML device should not be deemed defective if, on average, it exhibits a lower risk profile than an analogous device operated by humans, even if, on specific tasks, human operators still surpass the ML application.[96] There are two reasons for this. From a technical point of view, it will often be difficult or impossible to minimize ML error risks on all possible tasks simultaneously. More importantly, from a regulatory and societal point of view, a reduction of the overall risk level should be desirable, even if some specific risks (the stop sign) are more pronounced than with a human driver. This should be specified in a Level 1 regulation, and not simply left to the courts: ML developers will generally prefer clear regulatory guidance now to potentially slightly more tailored court rules a few years down the road.

This is not the end of the story, though. It would seem odd that a victim would be compensated if a human driver runs over a stop sign, but not if an autonomous vehicle makes the same mistake just because, on average, it is safer than a human driver. Granted, in some countries, there are rules according to which the person or organization deriving the economic benefits from a vehicle would be liable in such a case.[97] However, beyond autonomous vehicles, such rules may be lacking, for example if an otherwise supra-human medical AI tool fails to diagnose some specific cancer. Hence, if specific operational risks, particularly for life and limb, are higher than with a human agent, Level 2 regulation must ensure that the ML product may only be used with human oversight, even if, on average, it outperforms human actors.[98] A failure to manually override detectable machine error should then trigger fault-based liability of the human operator under Member State contract or tort law, solving the compensation problem.[99]

3. Post-market product monitoring and updates

Moreover, the possibility that the model may change continuously, explicitly by way of updates or implicitly through the analysis of new data (online learning[100]), must be considered.[101] In case of continuous product adaptations, post-market monitoring obligations must apply even after the product was first put into circulation. In some areas, they already form part of the existing legal requirements, for example concerning producer's liability under German tort law (which is independent of EU product liability),[102] and, partially,[103] in EU product safety law (post-market surveillance system).[104]

In this vein, product monitoring obligations should be added, at Level 1, to EU product liability law as well, which, so far, does not feature any such post-market obligations (cf. Art. 7(b) PLD),[105] and to the remaining areas of product safety law. Ultimately, updates and learning processes that significantly change product risks must be subject, in product liability and safety

[96] Cf. Mazzini (n 14) 17 et seq.; see also, for a lower risk boundary, Geistfeld (n 95) 1651-1654.
[97] See, e.g., § 7 StVG (German Road Traffic Act).
[98] Mazzini (n 14) 14 et seq.; an example is §1b StVG.
[99] See, e.g., Schönberger (n 91), 197; Hacker et al., 'Explainable AI under contract and tort law: legal incentives and technical challenges', 28 Artificial Intelligence and Law (2020), https://doi.org/10.1007/s10506-020-09260-6, at 10; Froomkin, Kerr and Pineau. 'When AIs outperform doctors', 61 Ariz. L. Rev. (2019), 33, 97 et seq.
[100] This can be avoided through offline learning, see van Wesel and Goodlo (n 69), 13.
[101] European Commission (n 4), 15.
[102] BGH, Case VI ZR 286/78, NJW 1981, 1606, 1607 et seq.
[103] See Mazzini (n 14) 8 et seq.
[104] See particularly Art. 10(10), Art. 83 of the Medical Devices Regulation (EU) 2017/745; furthermore Art. 5 of the General Product Safety Directive 2001/95/EC.
[105] See European Commission (n 4), 15; Mazzini (n 14), 22.

law, to the same conditions as the first placing on the market.[106] The latter can no longer be the only decisive point in time for testing and other legal duties of the producer; the 'later defect defence' of Art. 7(b) PLD must hence be modified.[107] Furthermore, a tort obligation must be recognized to not only monitor but also to actively update the product in case a safer software implementation becomes reasonably available to the producer.[108] In contract law, such a duty has now been installed by Article 8(2) in conjunction with recital 47 of the Directive on the supply of digital content and digital services[109] (particularly with respect to IT security). The exact contours of such an update obligation, however, remain to be specified at Levels 2-4.

4. Pre-market approval

A final question relates to the extent to which ML applications must be approved by competent authorities prior to putting them on the market, as discussed in the White Paper (prior conformity assessment).[110] Certainly, this cannot apply to all applications (Level 1), but only to those that pose a significant risk to important legal assets (Level 2).[111] In a first approach, it seems to make sense to answer this question by reference to the existing (product) approval law: If an approval procedure for traditional products is mandatory in one area (e.g. pharmaceuticals, medical devices, means of transport, machines), it should also be extended to ML applications in this area. Where this is not possible based on existing law, amendments are necessary.

Ensuring rigorous pre-market quality control seems particularly important in the medical and transport sectors, given the risks for life and limb. On the other hand, there is also a considerable public interest, e.g. in medical applications, in making effective ML instruments operational quickly and without unduly burdensome procedures, if this can lead to progress in medical care. The current Corona crisis vividly testifies to this urgency. In this respect, however, ML is not fundamentally different from other novel medical devices and products. Overall, this suggests a (cautious) transfer of the existing authorisation framework, which should also be reviewed for its effectiveness. In particular, a fast-track approval procedure for ML seems worth considering.

IV. Enforcement

A fourth focus should be on enforcement, which poses particular challenges given the technical complexity and lack of transparency of some ML models.[112]

1. General Dimensions

Generally speaking, enforcement will have to address issues concerning the proof of elements of liability ex post (a) and a regime for ex ante auditing to prevent damage in the first place (b).

---

[106] Expert Group on Liability and New Technologies – New Technologies Formation, Liability for Artificial Intelligence and Other Emerging Digital Technologies, 2019, 43.
[107] European Commission (n 4), 7 et seq.; Mazzini (n 14), 22 et seq.
[108] Geistfeld (n 95), 1646 et seq.; Wiebe (n 91), 630.
[109] Directive (EU) 2019/770.
[110] European Commission (n 2), 23 et seq.
[111] Even more lenient Scherer (n 16), 394: only voluntary certification with liability advantages.
[112] See, e.g., Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms', 3(1) Big Data & Society (2016), 1.

a) Linking substantial regulation and enforcement: Rebuttable presumptions, not strict liability

To counter the difficulties in proving, for example, defects and causality in ML applications, some commentators propose a strict liability regime for ML applications, potentially combined with a mandatory insurance regime.[113] However, in my view, it seems preferable to link the regulatory framework, particularly concerning documentation and training data, with liability rules by way of rebuttable presumptions and rules on the burden of proof.[114] This may strike a more equitable balance between the compensation interests of victims and the interest of society at large in ML innovation. Particularly for start-ups and SMEs, a strict liability regime and the concomitant insurance premia run the risk of being quite a substantial regulatory burden. By contrast, if, in the absence of a strict liability regime, ML developers are shielded from liability if they adhere to state-of-the-art techniques or (still to be developed) techno-legal standards (Level 3), this combines significant incentives with lower regulatory costs for developers. Strict liability therefore seems premature at this stage.[115]

Hence, to address enforcement difficulties, a breach of the aforementioned regulatory requirements for the creation of an ML model, in particular for documentation and training data, should trigger a rebuttable presumption (i) of a breach of duty on the part of the developer, and (ii) of the causality of this breach for any damage. Such a rule would, for example, greatly facilitate enforcement in product liability law and is also suggested in the White Paper and the Commission's Liability Report.[116] It would also enable redress of the producers, who may be different from the developers, against ML developers and therefore strengthen the incentives for careful ML development. The presumption could be rebutted, for example, by rigorous documentation and testing of the concerned model.

b) Auditing

This brings us to the question of auditing. The design of auditing procedures could also be left to national law. In principle, audits appear to be increasingly necessary to the extent that AI enters key economic and social areas. Other regulatory areas of relevance to society as a whole, such as tax law, social law and antitrust law, have long introduced audits and sector inquiries. Such a regime must therefore urgently be developed for the ML sector as well, and be reconciled with the legitimate business and innovation interests of developers and companies.

In the area of data protection law, audits are already provided for in Article 58(1)b GDPR. However, as mentioned, it is often unclear to what extent data used for ML training purposes are personal data and therefore the GDPR is applicable in the first place.[117] For this reason alone, the GDPR does not suffice. Second, the data protection authorities, already often under-resourced, should not be burdened with policing the entire ML sector, too.

It therefore seems sensible to set up a supervisory authority at EU or federal national levels, bundling both the legal and technical competencies in the ML area.[118] Audits could then be

[113] See, e.g., Čerka, Grigienė and Sirbikytė. 'Liability for damages caused by artificial intelligence', 31 Computer Law & Security Review (2015), 376, 386; Doshi-Velez et al., 'Accountability of AI under the law', arXiv preprint arXiv:1711.01134, at 5 and 11; Zech (n 91), 216 et seq.
[114] See also Zech (n 91), 218.
[115] Same result in Wagner (n 91), 14.
[116] European Commission (n 2), 15; European Commission (n 4), 15.
[117] Ostveen (n 29).
[118] Sachverständigenrat für Verbraucherfragen, Verbraucherrecht 2.0, 2016, 69 et seqq.; cf. also German Data Ethics Commission (n 10), 198.

carried out on a risk-based basis, by means of random checks as well as based on initial suspicions. Financial and human resources in line with these competences would be quite necessary. In terms of content, audits should cover the entire horizontal and, where appropriate, vertical regulatory ML law, with the joint involvement of specialist authorities (e.g., of the medical, automotive etc. sector) as necessary. If, however, the establishment of such a novel supervisory authority does not appear realistic in the foreseeable future, audits could also be carried out by independent, accredited, private auditing entities. Financial auditing can serve as a model here. Rules on conflicts of interest are then indispensable.

2. Specific enforcement of anti-discrimination law

Special enforcement tools should also be installed in anti-discrimination law, as a Level 2 regulation, where the lack of enforcement is particularly virulent at the moment.[119]

a) Access rights

In view of the manifest risk of discrimination when using ML, affected persons should be provided with an (ML-specific) right of the to access information on the statistical distribution of the output of the ML model between various protected groups (e.g. score distribution between genders). Only in this way can they determine whether there is any statistical difference at all and thus, possibly, (indirect) discrimination.[120] In the case of automated individual decision-making (Art. 22 GDPR), such an access right can, in my view, already be derived from Art. 15(1)(h) GDPR (information on the 'significance and the envisaged consequences' of the automated decision).[121] However, automated decisions within the meaning of Article 22(1) GDPR are quite rare because a human being typically intervenes at some point ('solely').[122] While Art. 15(1)(h) GDPR might also apply outside of automated decision-making processes ('at least in these cases'), this is fraught with considerable legal uncertainty. Moreover, in *Meister*, the CJEU explicitly rejected access rights, based on anti-discrimination law, in cases of suspected discrimination.[123]

Overall, therefore, the legal position of potential victims appears too weak. Hence, there is a need for legislative action at the European level introducing access rights concerning the statistical distribution of ML output between protected groups – aggregate data which should not count as trade secrets.

b) Public enforcement and collective redress

A reform of anti-discrimination enforcement, which need not be limited the ML, would furthermore have to include a public enforcement component and collective enforcement instruments (class action).[124] The latter has already been made optional in various anti-

[119] Chopin and Germaine, 'A comparative analysis of non-discrimination law in Europe 2015' (Report for DG Justice and Consumers, 2016), at 81 et seq.; Ellis and Watson, *EU Anti-Discrimination Law*, 2nd ed., 2012, 506; Craig and de Búrca, *EU Law*, 6th ed., 2015, 955 et seqq.
[120] See on this requirement CJEU, Case C-127/92, *Enderby*, para. 19; Case C-109/88, *Danfoss*, para.16.
[121] Hacker (n 30), 1173 et seq.
[122] Wachter, Mittelstadt and Floridi (n 60), 92; differently (substantial human influence needed to disapply Art. 22) Kamarinou, Millard and Singh. 'Machine learning with personal data', Queen Mary School of Law Legal Studies Research Paper 247/2016, 11 et seq.; Bygrave, 'Automated Profiling, Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling', 17 Computer Law & Security Report (2001), 17, 20.
[123] Case C-415/10, *Meister*, para. 46 et seq.; see also the discussion in Hacker (n 30), 1169 et seq.
[124] Hacker (n 30), 1170 et seq.; Martini (n 45), 310 f.

discrimination directives,[125] but has not been sufficiently implemented in all Member States (see, e.g., § 23 German Anti-Discrimination Act (AGG)).

c) Another rebuttable presumption

Anti-discrimination directives already contain alleviations of the burden of proof for victims.[126] These should be supplemented by an important aspect. In my view, the question of the justification of algorithmic discrimination chiefly depends on whether the discriminatory result of the ML model is based on biased modelling (e.g. training data containing historical bias) or on a statistically correct but unequal distribution of the values of the target variable between the protected groups (unequal ground truth).[127] However, this distinction, between biased modelling and unequal reality, cannot be made by potential victims without access to the data and the model.

Therefore, it should be clarified in an amendment to the evidentiary rules just mentioned that, where there is sufficient evidence that protected groups are disadvantaged, the user of the ML model must demonstrate that this is not based on biased modelling. Such proof could be provided, for example, by publishing metadata concerning the training data set and documenting the creation of the ML model (see above). If the user cannot prove that the disadvantage is not an artefact of modelling, a justification should in principle be denied.

V. Right to a data-free option

One final technological development that should be brought granted even more prominence in the AI agenda at the European level is the increasing combination of tracking technologies with models of machine learning and interconnected everyday objects to create an emerging *Internet of Everything.*[128] This suggests that, in the foreseeable future, contact with networked and machine-learning structures in both private and public spaces will be almost inevitable. To the extent that spaces devoid of data processing are shrinking, it becomes ever more important to legally guarantee, to the extent possible, the autonomous shaping of these networked spaces by the affected individuals. What seems necessary therefore is the establishment of a 'right to a data-free option' concerning data processing devices, including ML applications.[129] If chosen by the data subject, the service provider may then only process data strictly necessary for technical or legal reasons. Such a right must also be enforceable by digital means, for example by autonomous, personalized privacy assistants.[130] Data subjects would then be able to decide for themselves (or have assistants decide for them) in which areas they want to pay with data (data-intense option) and in which sensitive areas they would prefer to pay with conventional

[125] See for example Art. 20 of the Directive 2006/54.
[126] See, e.g., Art. 8 of the Race Equality Directive 2000/43/EC; Art. 10 of the Framework Directive 2000/78/EC; and Art. 9 of the Goods and Services Directive 2004/113/EC.
[127] Hacker (n 30), 1146 et seqq. on the technical distinction, 1160 et seqq. on the consequence for justification, esp. 1163 et seqq.
[128] See DeNardis, *The Internet in Everything*, 2020, 3 et seqq.; Breiner, Sriram and Subrahmanian, 'Compositional Models for the Internet of Everything', AAAI Spring Symposium Series (2018), 107.
[129] On this, see, e.g., Novotny and Spiekermann, 'Personal Information Markets AND Privacy: A New Model to Solve the Controversy', in: Hildebrandt et al. (eds.), *Digital Enlightenment Yearbook 2013*, 2013, 102, 107 et seq.; Strandburg, 'Free Fall: The Online Market's Consumer Preference Disconnect', University of Chicago Legal Forum (2013), 95, 170; Hacker and Petkova, 'Reining in the Big Promise of Big Data', 15 Northwestern Journal of Technology and Intellectual Property (2017), 1, 20 et seqq.; Becker, 'Reconciliating Data Privacy and Personal Data – A Right to Data-avoiding Products', 9 ZGE/IPJ (2017), 371.
[130] See, e.g., Das et al. 'Personalized privacy assistants for the internet of things: providing users with notice and choice', 17(3) IEEE Pervasive Computing (2018), 35.

money (data-free option). This seems paramount to safeguard data sovereignty in our increasingly interconnected and automated world.

E. Summary

The regulation of machine learning techniques is complex, but a worthwhile endeavour. It can only be successful if it is based on a rigorous analysis of existing law while also considering market solutions and risks to innovation. In terms of scope, such a regulation should generally be limited to machine learning and, in specific cases, be extended to software in general (product liability). To implement an ML regulation, this contribution suggests a multi-level legislative system in the sense of an adapted Lamfalussy procedure, inspired by modern EU capital markets law. Overarching principles, a legal framework for training data, general product liability and safety aspects, and enforcement provisions ought to be regulated horizontally (Level 1). At Level 2, these rules can be concretised in a sector- and application-specific manner. Legal certainty can be provided by specific *safe harbors* at Level 3 and guidelines at Level 4.

Concerning substantive regulation, the focus should be on five key projects: (1) the establishment of documentation requirements and access rules for the development of ML models; (2) a regulatory framework for ML training data that is independent of data protection law; (4) a reform of product liability and product safety law, with an extension of the former to software; (3) an improvement of law enforcement, including algorithmic auditing, but excluding strict liability for ML; and (5) the guarantee of a right to a data-free option. The particular challenge of any such regime will be to combine openness to technical development with legal certainty in such a way that innovations can succeed within a risk-adequate regulatory framework. The structure proposed here, consisting of principles, sector-specific rules, and safe harbors, seeks to bring these diverse objectives together.

A Legal Framework for AI Training Data

Author: Dr. Philipp Hacker, LL.M. (Yale)*

Building on the recently published White Paper of the EU Commission on Artificial Intelligence (AI), this article shows that training data for AI do not only play a key role in the development of AI applications, but are currently only inadequately captured by EU law. In this, I focus on three central risks of AI training data: risks of data quality, discrimination and innovation. Existing EU law, with the new copyright exception for text and data mining, only addresses a part of this risk profile adequately. Therefore, the article develops the foundations for a discrimination-sensitive quality regime for data sets and AI training, which emancipates itself from the controversial question of the applicability of data protection law to AI training data. Furthermore, it spells out concrete guidelines for the re-use of personal data for AI training purposes under the GDPR. Ultimately, the legislative and interpretive task rests in striking an appropriate balance between individual protection and the promotion of innovation. The law ought to provide support for the dynamic development of new AI models, but must also, where necessary, shape them in a socially desirable manner.

Keywords: artificial intelligence; training data; data protection law; anti-discrimination law; contract law; product liability; TDM exception; Commission White Paper on Artificial Intelligence

* AXA Postdoctoral Fellow, Faculty of Laws, Humboldt University of Berlin.

Contents

I. Problem definition and relevance

The use of Artificial Intelligence (AI) penetrates ever more areas of life. Therefore, it undoubtedly represents one of the great challenges of our time, both in economic and regulatory terms. This is demonstrated not least by the fact that the EU Commission has recently published a 'White Paper on Artificial Intelligence', which will form the basis for specific regulation of techniques and applications of AI at the EU level.[1] On multiple occasions, however, both the White Paper[2] and the accompanying Commission report on the liability and security of AI[3] mention an area at the intersection of law and AI which, so far, has hardly been analyzed from a legal perspective, and which is the focus of this paper: the regulation of data and data sets used for training AI applications.

1. AI training data: Technical background

Training data is, from a technical perspective, of fundamental importance for the development of AI applications. The techniques underlying AI can be roughly divided into three classes according to the type of learning strategies used:[4] supervised learning, unsupervised learning and reinforcement learning. Training data is the basis for both supervised learning and simulation environments in the area of reinforcement learning. These two techniques, in turn, are the basis for most of the AI applications currently in use, from automated face recognition,[5] credit scoring[6] and AI recruiting[7] to supra-human performance of AI agents in a number of complex games.[8]

In supervised learning, the algorithmic model is calibrated by matching predictions with (supposedly) correct results already contained in the training data.[9] In reinforcement learning, on the other hand, the AI develops an optimal strategy, which is determined on the basis of a learning environment consisting of data that sends feedback (*reward signals*) to the model.[10]

---

[1] European Commission, On Artificial Intelligence - A European approach to excellence and trust, White Paper, COM(2020) 65 final.
[2] European Commission, op. cit. *supra* note 1, p. 15, 18 et seq.
[3] European Commission, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64 final, 8 et seq.
[4] Russell/Norvig, *Artificial Intelligence*, 3rd ed., 2010, 694 et seq.; Shalev-Shwartz and Ben-David, *Understanding Machine Learning*, 2014, 4 et seq.; Jordan and Mitchell, 'Machine learning: Trends, perspectives and prospects', 349 *Science* (2015), 255 (257 et seq.); Goodfellow, Bengio and Courville, *Deep Learning*, 2016, 102 et seq.; Sutton and Barto, *Reinforcement Learning*, 2nd ed., 2018, 2.
[5] Lawrence et al., 'Face recognition: A convolutional neural-network approach', 8 *IEEE Transactions on Neural Networks* (1997), 98; Y. Sun et al., 'Deepid3: Face recognition with very deep neural networks', Working Paper, 2015, https://arxiv.org/abs/1502.00873; Goodfellow, Bengio and Courville, op. cit. *supra* note 4, p. 23 et seq.
[6] Fuster et al., 'Predictably Unequal? The Effects of Machine Learning on Credit Markets', Working Paper, 2018, https://ssrn.com/abstract=3072038.
[7] Faliagka et al., 'On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV', 42 *Artificial Intelligence Review* (2014), 515; Schmid Mast et al., 'Social Sensing for Psychology: Automated Interpersonal Behavior Assessment', 24 *Current Directions in Psychological Science* (2015), 154; Campion et al., 'Initial Investigation Into Computer Scoring of Candidate Essays for Personnel Selection', 101 *Journal of Applied Psychology* (2016), 958; Cowgill, 'Bias and productivity in humans and algorithms: Theory and evidence from resumé screening', Working Paper, 2018.
[8] Brown and Sandholm, 'Superhuman AI for multiplayer poker', 365 *Science* (2019), 885; Silver et al., 'Mastering the game of Go with deep neural networks and tree search', 529 *Nature* (2016), 484; Mnih et al., 'Human-level control through deep reinforcement learning', 518 *Nature* (2015), 529.
[9] LeCun, Bengio and Hinton, 'Deep Learning', 521 *Nature* (2015), 436 (436 et seq.); Goodfellow, Bengio and Courville, op. cit. *supra* note 4, pp. 79 et seqq., 102.
[10] Sutton and Barto, op. cit. *supra* note 4, p. 6; Jordan and Mitchell, op. cit. *supra* note 4, p. 258.

Here, too, the data of the learning environment is therefore of central relevance;[11] in addition, reinforcement learning often uses models (e.g. deep neural networks) that were initially 'pre-trained' with strategies of supervised learning (*deep reinforcement learning*).[12] Given the promises and risks associated with AI, training data therefore represent a key regulatory problem for the algorithmic society.[13]

2. Analytical framework and roadmap

This central position of training data has, however, not yet been sufficiently reflected in legal discussions. While studies on legal issues regarding the results and applications of AI already fill entire volumes,[14] training data still represents comparatively *terra incognita* in legal research.[15] Its central importance for machine learning techniques, however, suggests that, contrary to a widespread view,[16] it is not so much a regulation of *algorithms* as a regulation of *data* that is required – in particular, of the AI training data. For reasons of scope alone, however, the present contribution cannot cover all regulatory problems that arise in the context of AI training data (e.g. IT security). It therefore focuses on three central, interlinked risks and the way they are addressed in EU or (harmonized) Members State law: data quality risks, discrimination risks and innovation risks. All three also feature prominently in the Commission White Paper.[17] While this contribution puts the focus on AI training data used by private entities, its findings can be easily transferred, *mutatis mutandis*, to public actors.

The article begins with an examination of the three mentioned risks of training data (II.). On this basis, the regulatory requirements in existing EU data protection, anti-discrimination, general liability and intellectual property law for addressing these risks are analyzed (III.) and evaluated (IV.). This paves the ground for a discussion of potential policy reforms in an attempt to develop a risk-sensitive legal framework for AI training data (V.). Section VI. concludes.

II. The basic structure: Three regulatory risks

The three risks examined in this paper are data quality risks (1.), discrimination risks (2.) and innovation risks (3.). Each of them poses separate regulatory problems, but, at the same time,

---

[11] van Wesel and Goodloe, 'Challenges in the Verification of Reinforcement Learning Algorithms', NASA/TM-2017-219628, 15; Sutton and Barto, op. cit. *supra* note 4, p. 2.

[12] Silver et al., op. cit. *supra* note 8, pp. 484 et seq.; Mnih et al, op. cit. *supra* note 8, pp. 529 et seq.; Sutton and Barto, op. cit. *supra* note 4, pp. 236, 475.

[13] Cf. only the information provided by the Federal Commissioner for Data Protection and Information Security, BT-Drucks. 19/9800, p. 73; Pasquale, 'Data-Informed Duties in AI Development', 119 *Columbia Law Review* (2019), 1917 (1919 et seq.).

[14] See, e.g., Woodrow and Pagallo (eds.) *Research Handbook on the Law of Artificial Intelligence*, 2018; Wischmeyer and Rademacher (eds.), *Regulating Artificial Intelligence*, 2020; Vladeck, 'Machines without principals: liability rules and artificial intelligence', 89 Washington Law Review (2014), 11; Barocas and Selbst, 'Big Data's Disparate Impact', 104 *California Law Review* (2016), 671; Calo, 'Singularity: AI and the Law', 41 *Seattle University Law Review* (2017), 1123; Geistfeld, 'A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation', 105 California Law Review (2017), 1611; Surden, 'Artificial Intelligence and Law: An Overview', 35 *Georgia State University Law Review* (2018), 1305; Wachter and Mittelstadt, 'A Right to Reasonable Inferences', *Columbia Business Law Review* (2019), 494.

[15] Exceptions are Pasquale, op. cit. *supra* note 13; Gerberding and Wagner, 'Qualitätssicherung für „Predictive Analytics" durch digitale Algorithmen', *Zeitschrift für Rechtspolitik* (2019), 116.

[16] See, e.g., Tutt, 'An FDA for Algorithms', 69 *Administrative Law Review* (2017), 83; Lodge and Mennicken, 'The importance of regulation of and by algorithm', in: Andrews et al. (eds.), Algorithmic Regulation, LSE Discussion Paper No. 85, 2017, 2; Sachverständigenrat für Verbraucherfragen [German Consumer Affairs Council], Verbraucherrecht 2.0, Report, 2016, pp. 60 et seqq., especially p. 67.

[17] European Commission, op. cit. *supra* note 1, p. 3 et seq., 14 et seq., 19.

they are sufficiently interconnected to require and justify a uniform approach in the context of this study.

1. Quality risks

Data quality risks are central to machine learning.[18] They have direct implications for supervised learning techniques because objectively incorrect training data (typically) leads to incorrect model predictions.[19] However, data quality is not limited to objective correctness, but must also include, for example, the timeliness and representativeness of the data.[20] The development of legally operationalizable quality criteria for training data is therefore frequently called for[21] and is a central desideratum of this contribution (see below, V.2.a)i.).

The situation is even more complex in the area of reinforcement learning because there is often a lack of objective standards for assessing the 'correctness' of the training environment. If, for example, a system for controlling an autonomous vehicle is confronted with various problem situations in a simulator,[22] these constellations will rarely be objectively incorrect. At best, they can be qualified as unlikely or unbalanced. The problem is thus transformed into one of the adequate selection of representative use situations that the system has to cope with.

2. Discrimination risks

Training data is also a major source of algorithmic discrimination.[23] This is demonstrated by real cases from the fields of face recognition,[24] AI recruiting[25] and personalized advertising,[26] to name but a few examples. Discrimination risks are partly linked to, or may be a consequence of, data quality risks if and to the extent that the data quality for a particular protected group is on average negatively affected.[27]

However, this link does not necessarily exist; rather, discrimination risks may arise independently of quality risks. Even if the data quality is the same with regard to the different protected groups, the lack of group balance in a data set (e.g., the underrepresentation of a

---

[18] Kotsiantis/Kanellopoulos/Pintelas, 'Data preprocessing for supervised leaning', 1 *International Journal of Computer Science* (2006), 111 (116); Pasquale, op. cit. *supra* note 13, p. 1920 et seqq.

[19] Sachverständigenrat für Verbraucherfragen, Verbrauchergerechtes Scoring, Report, 2018, p. 83, 146; Hoeren, 'Thesen zum Verhältnis von Big Data und Datenqualität', *MultiMedia und Recht* (2016), 8 (11).

[20] German Data Ethics Commission (Datenethikkommission), Opinion of the Data Ethics Commission, 2019, 52; Sachverständigenrat für Verbraucherfragen (Expert Council on Consumer Affairs), op. cit. *supra* note 19, p. 145.

[21] Deussen et al., 'Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 1: Quality Meta Model', DIN SPEC 92001-1, 2019, 17; Information Commissioner's Office and The Alan Turing Institute, Explaining decisions made with AI, Part 2, Guidance, 2019, 28, 89; Artificial Intelligence Strategy of the German Federal Government, BT-Drucks. 19/5880, 36, 39; Sachverständigenrat für Verbraucherfragen,, op. cit. *supra* note 19, pp. 130-132, 144-146.

[22] See on this Gallas et al., 'Simulation-based reinforcement learning for autonomous driving', *Proceedings of the 36th International Conference on Machine Learning* (2019), 1.

[23] See in detail Barocas and Selbst, op. cit. *supra* note 14, p. 680 et seqq.; Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law', 55 *Common Market Law Review* (2018), 1143 (1146 et seqq.) and the evidence in the references *infra* note 135.

[24] Buolamwini and Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Conference on Fairness, Accountability and Transparency in Machine Learning (FAT*)* (2018), 77.

[25] Lowry and Macpherson, 'A blot on the profession', 296 *British Medical Journal* (1988), 657; see also Reuters, 'Amazon ditched AI recruiting tool that favored men for technical jobs', The Guardian (October 11, 2018), https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

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

[27] Cf. Information Commissioner's Office, Big data, artificial intelligence, machine learning and data protection, Version 2.2., 2017, para. 94-96.

protected group, so-called *sampling bias*) can lead to systematic distortions and discrimination.[28] Nonetheless, it must be recognized that decisions made by humans can also be guided to a considerable extent by conscious or unconscious bias.[29] In contrast to human decisions, however, the parameters of machine models can be explicitly and directly regulated,[30] for which the computer science literature on discrimination-aware machine learning (*algorithmic fairness*) offers manifold starting points.[31]

3. Innovation risks

In a technical environment as dynamic as that of AI, however, innovation risks must also be considered. They are divided into two dimensions. First, an independent, innovation-relevant 'blocking risk' must be recognized. This is because data may be subject to intellectual property rights or may be protected by data protection laws; this, in turn, makes its use as training data considerably more difficult.

Second, there is an overarching risk of over-regulation, which may unduly inhibit the development of AI due to significant or even prohibitive costs for the addressees (regulatory cost risk). This is, however, first and foremost a question of calibrating the respective regulatory burden, which must be considered, in the following, in the individual legal requirements addressing the risks just mentioned.

III. Existing legal requirements for training data

The existing regulatory requirements can equally be broken down into norms addressing quality risks (1.), discrimination risks (2.) and the risk of blockage through intellectual property rights and data protection law (3.).

1. Quality risks

In extant EU law, risks concerning the quality of training data are partly covered by data protection law and, more indirectly, by general liability law, such as contract and tort law.

a) Data protection law

Although data protection law provides for a number of data quality requirements (i.), their effectiveness depends on the questionable applicability of data protection law to training data (ii.).

i. Requirements

Within its scope of application, the GDPR not only requires a legal basis for all processing of personal data, including for AI applications (Article 6(1) GDPR), but also contains some starting points for ensuring data quality.

---

[28] Calders and Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures', in: Custers et al. (eds.), *Discrimination and Privacy in the Information Society*, 2013, 43 (51); on sampling bias generally Hand, 'Classifier Technology and the Illusion of Progress', 21 *Statistical Science* (2006), 1 (8 et seq.).

[29] See only Greenwald and Krieger, 'Implicit Bias: Scientific Foundations', 94 *California Law Review* (2006), 945 (948 et seqq.).

[30] See, e.g., Kleinberg et al., 'Human decisions and machine predictions', 133 *The Quarterly Journal of Economics* (2018), 237 (242 et seq.).

[31] Overview in Dunkelau and Leuschel, 'Fairness-Aware Machine Learning', Working Paper, 2019.

(1) The accuracy principle, Art. 5(1)(d) GDPR

For example, the principle of accuracy laid down in Article 5(1)(d) GDPR stipulates that personal data must be 'accurate and, where necessary, kept up to date'. Data subjects have a corresponding right to have inaccurate data rectified, Article 16 GDPR. However, it is still largely unclear how the very general accuracy principle embodied in Article 5 GDPR can be legally operationalized for the area of training data.[32] This is crucial, however, as the violation of an Article 5 principle not only triggers liability according to Article 82 GDPR, but also fines of up to 4% of the global annual turnover according to Article 83(5) GDPR.

For example, in terms of accuracy, it will make a difference if, in a data set containing 100,000 data points, one data point is slightly inaccurate (e.g., yearly income of an individual registered as €50,000 instead of €51,000) or if a large number of data points are incorrect by a large margin.[33] While a slight inaccuracy of a single data point in the training data may not (significantly) change the resulting AI model,[34] such an error may be much more consequential if it concerns the input data of an individual actually analyzed by the model.[35]

The GDPR, however, does not specify any metric to measure accuracy; in fact, it does not even clearly state if the degree of accuracy makes a difference (i.e., the margin of error), or if 'inaccurate remains inaccurate', irrespective of how close the processed value is to the correct one. Similarly, it is not specified in which cases the data needs to be kept up to date, and with what frequency records must be updated. In a very general manner, Article 5(1)(d) GDPR only requires that controllers must take 'every reasonable step […] to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay'. Proposals to make this regime more concrete can be based on a broad literature from computer science dealing with data quality, but should ultimately be developed outside the GDPR (see below, III.1.a)ii.(2) and V.2.a)).

(2) Member State data protection law and the primacy of EU law

National data protection law, on the other hand, sometimes contains more specific provisions. For example, German data protection law provides for a specific regulation on scoring in § 31 of the German Data Protection Act (BDSG). According to § 31 BDSG, data used for scoring, and hence also training data, must be 'significant'. It may only be considered if this significance is derived by a 'scientifically recognized mathematical-statistical procedure' (§ 31(1) no. 2 BDSG). Furthermore, it is prohibited to base scoring exclusively on address data of the data subject (§ 31(1) no. 3 BDSG).

Although these regulations contain worthwhile points of reference for a regulatory framework for training data, they are, in their concrete form, quite problematic in several respects. First, the sweeping reference to 'recognized mathematical-statistical methods' is highly imprecise,

---

[32] Similar conclusion in Mitrou, 'Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) 'Artificial Intelligence-Proof'?', Working Paper, 2018, https://ssrn.com/abstract=3386914, 51 et seq.; Sachverständigenrat für Verbraucherfragen, op. cit. *supra* note 19, p. 131; Hoeren, 'Big Data und Datenqualität –ein Blick auf die DS-GVO', *Zeitschrift für Datenschutz* (2016), 459 (461 et seq.); Roßnagel, in: Simitis/Hornung/Spiecker gen. Döhmann (eds.), *Datenschutzrecht*, 2019, Art. 5 DSGVO para. 148 et seq.; approaches for a concretisation in Hoeren, op. cit. *supra* note 19, p. 8.

[33] Cf. Mitrou, op. cit. *supra* note 32, p. 52; Butterworth, 'The ICO and artificial intelligence: The role of fairness in the GDPR framework', 34 *Computer Law & Security Review* (2018), 257, 260 et seq.

[34] Mayer-Schönberger and Cukier, *Big Data*, 2013, 32 et seqq.

[35] Information Commissioner's Office, op. cit. *supra* note 27, para. 92.

since completely different mathematical operations underlying different machine learning methods are involved here, none of which, however, evidently represents a scientific gold standard.[36] Only arbitrariness and chance are therefore clearly excluded.

Second, and most importantly, it is currently debated to what extent § 31 BDSG is compatible with the primacy of EU law in the first place, as the GDPR does not contain an explicit opening clause for scoring.[37] Considering the structure of the GDPR, the application of § 31 BDSG is indeed precluded by the primacy of EU law. Admittedly, the GDPR does not contain a specific regime for scoring and may thus be (deliberately) under-complex. However, scoring is certainly not an issue that was not considered at all by the drafters of the GDPR and, implicitly, left to Member State law, as the provisions on profiling clearly show (e.g., Article 4(4), 22 GDPR). Hence, scoring is covered by the general concepts and requirements of, e.g., Article 5(1), Article 6(1)(f) GDPR. Some scholars claim that, because these standards are so vague, they contain an implicit mandate for national data protection law to render them more concrete.[38] However, this stands in clear contradiction with the specifically defined opening clauses and restriction possibilities for Member State law, for example in Article 23 and Article 89 GDPR. Therefore, outside of such opening clauses, a concretization of general GDPR standards, even if they only contain vague and undefined legal concepts, can only be carried out by the European legislator or the CJEU, but not unilaterally and in a potentially diverging manner by the Member States. Otherwise, the harmonizing effect of EU data protection law would be fully dissolved.[39]

ii. Applicability of the GDPR

All of these prerequisites, however, only apply if the GDPR (or national data protection acts like the BDSG) are applicable at all *ratione materiae*. The central precondition for this is that the training data must qualify as personal data in accordance with Article 4(1) GDPR. The decisive element is whether a natural person is directly or indirectly identifiable. The GDPR regime therefore excludes legal persons from the outset.[40] Furthermore, training data is often anonymized by removing directly identifying information (e.g. names) or applying more powerful de-identification techniques.[41] For this reason, it is sometimes assumed in the literature that training data tends not to fall under the regime of the GDPR.[42]

---

[36] See Sachverständigenrat für Verbraucherfragen, op. cit. *supra* note 19, pp. 131 et seq., 144; Domurath and Neubeck, 'Verbraucherscoring aus Sicht des Datenschutzrechts', Working Paper, 2018, 24; Gerberding and Wagner, op. cit. *supra* note 15, p. 118.

[37] Buchner, in: Kühling/Buchner (eds.), DS-GVO/BDSG, 2nd ed. 2018, § 31 BDSG para. 4 et seq.; Moos and Rothkegel, 'Nutzung von Scoring-Diensten im Online-Versandhandel', *Zeitschrift für Datenschutz* (2016), 561 (567 et seq.).

[38] Taeger, 'Scoring in Deutschland nach der EU-Datenschutzgrundverordnung', *Zeitschrift für Rechtspolitik* (2016), 72 (74).

[39] Similarly Moos and Rothkegel, op. cit. *supra* note 37, p. 567 et seq.; for the autonomous interpretation of EU law in general, see only CJEU, Case C-395/15, *Daouidi*, para. 50; Case C-673/17, *Planet49*, para. 47.

[40] But see CJEU, Joined Cases C-92/09 and C-93/09, *Schecke*, para. 52 et seq. on the applicability of Art. 7 and 8 of the Charter.

[41] For an overview of such techniques, see El Emam, Rodgers and Malin, 'Anonymising and sharing individual patient data', 350 *BMJ* (2015), h1139; Cavoukian and Castro, Big data and innovation, setting the record straight: de-identification does work, Office of the Information and Privacy Commissioner, Ontario, 2014, 9-11.

[42] Ostveen, 'Identifiability and the applicability of data protection to big data', 6 *International Data Privacy Law* (2016), 299, 307; see also Hintze, 'Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency', 8 *International Data Privacy Law* (2018), 86 (89).

(1) Re-identification strategies, *Breyer*, and illegality

However, as numerous empirical studies have shown,[43] data can, under certain conditions,[44] be effectively de-anonymized. Concerning training data, an indirect reference to a person can be established in two ways.[45] First, re-identification can take place when information containing a link between the data and specific data subjects has been removed from the data set, but can still be accessed by the controller or a third party (e.g., a list of real names linking them to unique identifiers in the data set).[46] Second, even if such a file does not exist, technical de-anonymization strategies can be executed on the basis of existing data and without recourse to directly identifying information.[47]

However, not every possibility of re-identification leads to identifiability in the sense of the Article 4(1) GDPR. In this respect, the CJEU decided in the landmark *Breyer* case, on the identical requirements of the 1995 Data Protection Directive, that it must be reasonably likely that the controller will use the strategies available to him to carry out the identification.[48] Given the wide variety of technical re-identification strategies, it may seem at first glance that large amounts of training data typically represent personal data, since the probability of re-identification usually increases with the amount of data.[49] However, on a technical level, this overlooks that actual re-identification is often much harder than the empirical studies proving certain attack strategies seem to imply, particularly when state-of-the-art de-identification techniques are used.[50] Furthermore, it has not been sufficiently taken into account by the legal literature that the CJEU categorically rejects a sufficient probability of indirect identification if the means of identification would be *illegal.*[51]

This, in turn, raises the as yet unresolved questions as to the extent to which (a) technical re-identification strategies would actually be illegal, e.g. due to a violation of Articles 5, 6 or 9 GDPR, and whether (b) such illegality would indeed categorically exclude any identifiability according to Article 4(1) GDPR. On the first question, the legality of re-identification will, most importantly, have to be measured against Article 6(1)(f) GDPR. The result will therefore crucially depend on whether the party conducting the de-anonymization may advance

---

[43] See, e.g., Sweeney, 'Uniqueness of Simple Demographics in the U.S. Population, Laboratory for International Data Privacy', Working Paper LIDAP-WP4, 2000; Narayanan and Shmatikov, 'Robust De-anonymization of Large Datasets', *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (2008), 111; Rocher, Hendrickx and de Montjoye, 'Estimating the success of re-identifications in incomplete datasets using generative models', 10 *Nature Communications* (2019), 3069.

[44] See the careful and cautionary analysis in Cavoukian and Castro, op. cit. *supra* note 41, 2-8.

[45] Information Commissioner's Office, op. cit. *supra* note 27, para. 132-136; Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymisation Techniques', WP 216, 2014, 8 et seq.

[46] Information Commissioner's Office, op. cit. *supra* note 27, para. 136.

[47] Overview at Article 29 Data Protection Working Party, Opinion 5/2014 on anonymisation techniques, WP 216, 2014, 13; Ohm, 'Broken Promises of Privacy', 57 *UCLA Law Review* (2009), 1701 (1723 et seq.).

[48] CJEU, Case C-582/14, *Breyer*, para. 45-49.

[49] See Ostveen, op. cit. *supra* note 42, p. 307; Veale, Binns and Edwards, 'Algorithms that remember: model inversion attacks and data protection law', 376 *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2018), Article 20180083, 6 et seq.

[50] El Emam, 'Is it safe to anonymize data?', The BMJ Opinion (February 6, 2015), http://blogs.bmj.com/bmj/2015/02/06/khaled-el-emam-is-it-safe-to-anonymize-data/; Cavoukian and Castro, op. cit. *supra* note 41, 2-8; El Emam et al., 'De-identification methods for open health data: the case of the Heritage Health Prize claims dataset', 14(1) *Journal of Medical Internet Research* (2012), e33; El Emam et al., 'A systematic review of re-identification attacks on health data', 6(12) *PloS one* (2011), e28071; see also Hintze, op. cit. *supra* note 42, at 90.

[51] CJEU, Case C-582/14, *Breyer*, para. 46; on the legality requirement specifically Kühling and Klar, 'Speicherung von IP-Adressen beim Besuch einer Internetseite', *Zeitschrift für Datenschutz* (2017), 27 (28).

compelling and legitimate interests. On the one end, fraud and crime prevention may justify such an act (recital 50 GDPR); on the other end, marketing purposes quite clearly should not, as this would directly contradict and defeat the purpose of anonymization.

This result, however, gives rise to the follow-up question of whether illegality must be considered, for the purposes of the *Breyer* analysis, in a concrete instance or merely in the abstract. In the former case, if there is no concrete fraud or crime suspicion against the specific data subject, the means of re-identifiability must be qualified as illegal. In the latter case, toward which the CJEU seems to lean,[52] re-identification strategies must count as legal as it can never be generally ruled out that, in some scenario, there would be legitimate interests justifying them. This consequence, however, clearly speaks against the abstract perspective: it would deprive the criterion of illegality of any meaning, as one can always imagine situations in which re-identification would be legal. Therefore, the legality requirement must be based on a concrete analysis at the moment at which a (potential) re-identification takes place. This suggests that, absent fraud or crime suspicions, de-anonymization will generally be illegal both under the GDPR and for the *Breyer* analysis. It is precisely the purpose of data protection law to guard against clandestine re-identification of individual persons, particularly in conjunction with potentially large data sets such as training data.

This makes the question of the consequences of the illegality of de-anonymization all the more virulent. The case law of the CJEU seems to suggest an easy answer: illegality excludes identifiability.[53] This would, however, have the perplexing consequence that, if identified by illegal means, the persons concerned would not enjoy the protection of the GDPR, while it would have these benefits if it was identified in a legal way. From a teleological point of view, this is not convincing, as the protective regime of the GDPR seems, if anything, more important in case of illegal identification. This conclusion is supported by the fact that recital 26 of the GDPR, in discussing identifiability, does not mention the legality criterion. On this reading, then, even the possibility of illegal re-identification must be factored into the risk analysis. On the other hand, not every remote, potentially illegal means of re-identification may suffice to establish identifiability, since otherwise anonymization would be virtually impossible.[54] This would not only dramatically reduce incentives to employ de-identification in the first place,[55] but also contravene the spirit of recital 26 GDPR, which clearly presupposes that the exclusion of the applicability of the GDPR, by means of anonymization, must be possible.[56] Therefore, a risk-based approach, which is generally followed in the GDPR,[57] ought to be pursued, which puts the data protection-specific risks (see only recital 75 of the GDPR) in relation to the risk of re-identification, even by illegal means.[58] Under this understanding, only a concrete re-

[52] CJEU, Case C-582/14, *Breyer*, para. 47 et seq.; see also Finck and Pallas, 'They Who Must Not Be Identified - Distinguishing Personal from Non-Personal Data Under the GDPR', International Data Privacy Law (forthcoming), https://ssrn.com/abstract=3462948, 14.
[53] CJEU, Case C-582/14, *Breyer*, para. 46; Purtova, 'The law of everything. Broad concept of personal data and future of EU data protection law', 10 *Law, Innovation and Technology* (2018), 40 (64).
[54] Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymisation Techniques', WP 216, 2014, 5; Karg, in: Simitis/Hornung/Spiecker gen. Döhmann (eds.), Datenschutzrecht, 2019, Art. 4 Nr. 1 DS-GVO para. 64; Brink and Eckhardt, 'Wann ist ein Datum ein personenbezogenes Datum?', *Zeitschrift für Datenschutz* (2015), 205 (211).
[55] Cf. Hintze, op. cit. *supra* note 42.
[56] Cf. Information Commissioner's Office, op. cit. *supra* note 27, para. 130; Finck and Pallas, op. cit. *supra* note 52, p. 15.
[57] See, e.g., Lynskey, *The Foundations of EU Data Protection Law*, 2015, 81 et seqq.; Article 29 Data Protection Working Party, 'Statement on the role of a risk-based approach in data protection legal frameworks', WP 218, 2014, 2; *Gellert*, 'Data protection: a risk regulation?', 5 International Data Privacy Law 2015, 3.
[58] Cf. Purtova, op. cit. *supra* note 53, p. 64 et seq.; Information Commissioner's Office, op. cit. *supra* note 27, para. 134 et seq.; implicitly also Finck and Pallas, op. cit. *supra* note 52, p. 15 et seq.

identification risk that is reasonably likely and normatively sufficiently relevant triggers the applicability of the GDPR.[59]

(2) Conclusions for supervised and reinforcement learning

The consequence of this analysis, however, is that strong anonymization strategies,[60] unless there is evidence of a concrete (legal or illegal) re-identification intention, tend to exclude the applicability of the data protection regime to training data. Even with training data used for supervised learning, the applicability of the GDPR is therefore highly questionable and will often have to be denied.[61] This holds even more in the case of the training environments for reinforcement learning, which, as far as can be seen, have not been considered by legal analysis at all so far. If simulation environments operate with hypothetical scenarios (synthetic data),[62] a reference to identifiable, real persons will be completely excluded. Nevertheless, questions of the quality of this training environment remain highly relevant to the results of the learning process.[63]

In sum, it therefore does not seem appropriate to make the regulatory framework for training data depend on whether the threshold for identifiability has just been exceeded or not yet. The view must therefore be broadened beyond data protection law.

b) General liability law

In this endeavor, general liability law seems an obvious candidate. It can, in principle, contribute to the internalization of technological risks.[64] However, the application requirements and substantive standards for quality risks of training data need to be examined more closely in this regime, too.

i. Contract Law

Not much research has been devoted yet to the question of the extent to which poor training data quality may constitute a non-conformity of the trained product that is relevant under contract law.[65] Insofar as high-quality training data, in individual cases, represent a contractual condition at all, non-conformity under a sales contract (Art. 2 of the Consumer Sales Directive),[66] a rental or a service contract comes into consideration here. Furthermore, liability may also be based on (the implementation of) Art. 11 et seq. of the Directive on Digital Content

[59] Similar result in Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymisation Techniques', WP 216, 2014, 6 et seq., 10, without, however, the discussion of illegal re-identification.
[60] On workable strategies, such as randomization and generalization, see, e.g., Cavoukian and Castro, op. cit. *supra* note 41, 9-11; Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymisation Techniques', WP 216, 2014, 11 et seqq.; and the references *supra* note 50.
[61] See Winter, Battis and Halvani, 'Herausforderungen für die Anonymisierung von Daten', *Zeitschrift für Datenschutz* (2019), 489 (490, 492).
[62] See Gallas et al., op. cit. *supra* note 22.
[63] See the references *supra* note 11.
[64] Jacob and Spaeter, 'Large-Scale Risks and Technological Change', 18 *Journal of Public Economic Theory* (2016), 125 (126 et seq.).
[65] Very brief remarks in Schuhmacher and Fatalin, 'Compliance-Anforderungen an Hersteller autonomer Software-Agenten', *Computer und Recht* (2019), 200 (203 et seq.); on liability for IT security defects, see, e.g., Pinkney, 'Putting blame where blame is due: Software manufacturer and customer liability for security-related software failure', 13 *Alb. LJ Sci. & Tech.* (2002), 43 (69 et seqq.); Raue, 'Haftung für unsichere Software', *NJW* (2017), 1841.
[66] Directive 1999/44/EC.

and Digital Services (DCDS Directive),[67] depending on the type of contract governing the AI application. However, unless a training data set is itself the object of the transaction, it will generally be the undesirable properties of the trained product, and not the quality defects of the training data, which will be considered the non-conforming feature vis-à-vis the end-user.

Furthermore, if data protection law *is* applicable, a possible violation of the GDPR, as a result of the quality deficit (Art. 5(1)(d) GDPR), may constitute a contractually relevant defect in an AI application. This is also suggested by recital 48 of the DCDS Directive.[68] The inclusion of data protection requirements in the contractual target quality of an AI application can be subjectively agreed upon (Art. 2(2)(a) and (b) of the Consumer Sales Directive); or it may objectively fall under the general fit-for-purpose provision or the reasonable quality expectations of the buyer (Art. 2(2)(c) and (d) of the Consumer Sales Directive).[69] For example, a personalized privacy assistant, supposed to help the end-user navigate privacy choices,[70] would arguably be held to be in breach of contractual conformity requirements if it did not comply with the GDPR, including if it was calibrated on faulty training data and therefore violated the discussed GDPR data quality standards.

Finally, contractual interpretation will often suggest that, for products or services closely linked to data processing, at least the compliance with the essential requirements of the GDPR will constitute an ancillary contractual obligation.[71] The German Federal Court of Justice (BGH), in a landmark case, ruled that the basic requirements of the public law regime of securities regulation, rooted in EU law, generally do form part of the contractual obligations in an investment advice contract.[72] The spirit of this ruling could and should be transferred to the relationship between EU data protection and national contract law. Such additional contractual liability rules are relevant in spite of the existence of Article 82 GDPR as the counterparty of the end-user need not be identical with the data controller liable under the GDPR.

This finding, however, immediately points to an incentive problem. There will often be no direct contractual relationship between the developer of an AI application and the end-user, i.e. the injured party. For this reason, incentives for those developers who handle the training data can only arise, under contract law, through redress along the sales/contract chain (e.g. according to Art. 4 Consumer Sales Directive, Art. 20 DCDS Directive). This will be of importance for the evaluation of existing data quality law (below, V.2.a)iii.(1)).

ii. Tort law

Beyond contract law, it is quite conceivable that a quality deficiency of the training data, which manifests itself in an erroneous prediction of the algorithmic model, could also amount to a defect in the sense of Article 1 of the Product Liability Directive.[73] However, it is already highly questionable whether AI applications fall under the concept of product (within the meaning of

[67] Directive (EU) 2019/77.
[68] In this sense also Sein and Spindler, 'The new Directive on Contracts for Supply of Digital Content and Digital Services–Conformity Criteria, Remedies and Modifications–Part 2' 15 *European Review of Contract Law* (2019), 365 (371 et seq.).
[69] Cf. Faust, in: Beck'scher Onlinekommentar, BGB, 52nd ed. 2019, § 434 para. 68 (on product safety law violations as a contractual non-conformity).
[70] See, e.g., Das et al. 'Personalized privacy assistants for the internet of things: providing users with notice and choice', 17(3) *IEEE Pervasive Computing* (2018), 35.
[71] Cf. Gola and Piltz, in: Gola (ed.), DS-GVO, 2nd ed. 2018, Art. 82 para. 21.
[72] Bundesgerichtshof, Case XI ZR 147/12, *NJW* 2014, 2947 para. 36 f.
[73] Schuhmacher and Fatalin, op. cit. *supra* note 65, p. 204; see also Zech, 'Künstliche Intelligenz und Haftungsfragen', *Zeitschrift für die gesamte Privatrechtswissenschaft* (2019), 198 (209).

Art. 2 of the Product Liability Directive),[74] as they are typically at least also, if not primarily, intangible objects (software) and may contain service elements.[75]

According to Article 4 of the Product Liability Directive, the plaintiff must prove the damage, the defect and the causal link between the two. However, since internal processes of the producer are generally beyond the reach of the injured party, jurisprudence has reacted with significant alleviations to fulfil the burden of proof.[76] It would not be justified to withhold these benefits to parties injured by traditional software or AI applications. The development risks and the difficulties of plaintiffs in proving defects do not differ significantly between traditional products and software, including AI applications. If anything, the complexity and intransparency of AI models[77] make it even harder to trace damages to development defects.[78] The Commission[79] and the Expert Group on Liability and New Technologies[80] are therefore right to consider extending product liability (and product safety) law to AI applications in this respect in the future. Ultimately, all software, not only AI applications, should be covered.[81]

At the moment, however, this is the preferable, but a highly uncertain interpretation of product liability law. Furthermore, the incentive effect of this branch of law is also limited by the fact that a claim is restricted to cases of personal bodily injury and damage to privately used property (Art. 9 of the Product Liability Directive).[82] Therefore, product liability may become relevant when physically embodied robots are used, but it does not cover cases in which the algorithmic model provides predictions that lead to a merely pecuniary damage (e.g., credit scoring).

Finally, it should be borne in mind that determining the producer of a product (Art. 3 of the Product Liability Directive) can also pose considerable difficulties due to the cooperation practices customary in the IT industry concerning the development of code and the exchange of training data.[83] Overall, this results in a picture of liability law which is comparable to that of data protection law: both the conditions for application and the substantive standards for addressing quality risks in training data are subject to significant legal uncertainty.[84] We shall return to this issue below (V.2.a)iii.).

---

[74] See Schönberger, 'Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications', 27 *International Journal of Law and Information Technology* (2019), 171 (198 et seq.); Wagner, 'Robot Liability', Working Paper, 2018, https://ssrn.com/abstract=3198764, 11.
[75] Cf. CJEU, Case C-495/10, *Dutrueux*, para. 39: services providers not covered by the Product Liability Directive.
[76] See CJEU, Case C-621/15, *Sanofi Pasteur*, para. 43 (discussing evidentiary rules in French law); for German law, see, e.g., Wagner, in: Münchener Kommentar, ProdHG, 7th edition 2017, § 1 para. 72 et seqq.
[77] Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms', 3(1) *Big Data & Society* (2016), 1.
[78] Gurney, 'Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles', U. Ill. J. L. & T. (2013), 247 (265 et seq.).
[79] European Commission, op. cit. *supra* note 1, p. 14, 16; European Commission, op. cit. *supra* note 3, p. 14.
[80] Expert Group on Liability and New Technologies – New Technologies Formation, Liability for Artificial Intelligence and Other Emerging Digital Technologies, 2019, 42 et seq.
[81] For an analogous application of Art. 2 Product Liability Directive to software Wagner, op. cit. *supra* note 74; Zech, op. cit. *supra* note 73, p. 212; against this Schönberger, op. cit. *supra* note 74, p. 199.
[82] National tort law may, however, go beyond that, see Art. 13 of the Product Liability Directive.
[83] Zech, 'Risiken digitaler Systeme', Weizenbaum Series #2, 2020, 33 et seqq.; see also European Commission, op. cit. *supra* note 3, p. 13 et seq.; Günther, *Roboter und rechtliche Verantwortung*, 2016, 172 et seq.
[84] Same result in Hoeren, op. cit. *supra* note 19, p. 9.

2. Risk of discrimination

The second risk identified in this paper in connection with training data is that of discrimination against legally protected groups. This risk is primarily addressed by the anti-discrimination law, but also in part by data protection and general liability law.

a) Anti-discrimination law

The scope of EU anti-discrimination, however, law does not directly cover the compilation of training data or training environments. This is because such preparatory activity itself does not fall under any of the categories determining the material scope of the anti-discrimination directives, such as employment, social protection and advantages, access to publicly available goods and services, or education.[85] However, if an imbalance in the training data leads to results of the algorithmic model (output) that significantly disadvantage legally protected groups in any of these enumerated areas, this will usually amount to legally relevant (indirect) discrimination.[86] Therefore, liability for discrimination, enshrined in Member State law, could in principle provide incentives to design training data and processes in a way that mitigates discrimination.

However, it appears problematic that the enforcement of anti-discrimination law, which is left almost entirely to the initiative of the injured party, has considerable deficits.[87] Not only is there a significant risk of litigation, because in view of the partly supra-human performance of AI models, discrimination may be justified.[88] In particular, potentially injured parties will typically not even be able to prove a prima facie case of statistical inequality in the treatment of the different groups, which is necessary for indirect discrimination.[89] To achieve this, they would need access to the training data and the algorithmic model. However, the CJEU ruled in the *Meister* case that the mere presumption of discrimination is not a ground for recognizing rights of access.[90] In the absence of effective enforcement opportunities, liability arising from a breach of anti-discrimination provisions therefore has but a small incentive effect, also at the training data stage.

b) Data protection, contract and tort law

This enforcement deficit could be reduced if algorithmic discrimination also constituted a violation of data protection law and its enforcement instruments (Art. 82 et seqq. GDPR), which have been considerably strengthened in the GDPR, could be used. It could certainly be argued that algorithmic discrimination is also relevant in terms of data protection law, both for the principles of fair and accurate data processing (Article 5(1)(a) and (d) GDPR) and for automated individual decision making (Article 22(3) GDPR).[91] However, here again the problem arises

---

[85] See, e.g., Art. 3(1) of the Race Equality Directive 2000/43/EC; Art. 3(1) of the Framework Directive 2000/78/EC; and Art. 3(1) of the Goods and Services Directive 2004/113/EC.

[86] Hacker, op. cit. *supra* note 23, p. 1151 et seqq.

[87] Chopin and Germaine, 'A comparative analysis of non-discrimination law in Europe 2015' (Report for DG Justice and Consumers, 2016), at 81 et seq.; Ellis and Watson, *EU Anti-Discrimination Law*, 2nd ed. (OUP, 2012), 506; Craig and de Búrca, *EU Law*, 6th ed. (OUP, 2015), 955 et seqq.

[88] More precisely Hacker, op. cit. *supra* note 23, p. 1160 et seqq.; Schönberger, op. cit. *supra* note 74, p. 184 et seq.

[89] See on this requirement CJEU, Case C-127/92, Enderby, para. 19; Case C-109/88, Danfoss, para.16; see also references in note 88.

[90] CJEU, Case C-415/10, Masters, para. 46.

[91] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679', WP 251 rev. 1, 2018, 10, 14, 27 et seq.; Hacker, op. cit. *supra* note 23, p. 1171 et seq.

that training data itself may not constitute personal data and the applicability of the GDPR may therefore be excluded.

With regard to contractual and tort liability law, on the other hand, the findings are similar to those for quality risks: although algorithmic discrimination may constitute a non-conforming feature under contract law or a defect under product liability law,[92] the fulfilment of the other liability requirements (existence of a contract; product etc.) is, as seen, cast into serious doubt. Overall, therefore, the risk of discrimination arising from unbalanced training data does not seem to be adequately addressed by existing anti-discrimination, data protection and general liability law.

3. Blocking risk

The third risk relevant to the requirements for training data stems from the raw data used for AI training purposes. Crucially, this data can be covered by data protection law (a) or even by intellectual property rights (b), imposing potentially severe constraints on the re-use of that data for AI training purposes. This has recently also been recognized by the EU Commission in its Communication on a European data strategy.[93]

a) Data protection law

In Article 6(4) of the GDPR, EU data protection law sets out, as an expression of the purpose limitation principle (Article 5(1)(b) GDPR), specific requirements for changing the purpose of personal data. They apply, for example, if data originally collected with a different aim is now supposed to be used as training data. Given the tension between the second sentence of recital 50 of the GDPR, which maintains that the re-use can be based on the legal ground for collecting the data in the first place, and the structure of Article 6, with the requirements of Article 6(1) GDPR applying in principle to every new data processing step, there is quite some debate as to whether Article 6(4) GDPR constitutes a separate legal base for processing, besides Article 6(1) GDPR.[94] Both from a systematic and from a teleological perspective, this view must be rejected: the requirements in Article 6(4) GDPR are arguably less strict than those in Article 6(1)(f) GDPR, which would lead to the untenable conclusion that it is easier to use personal data for a secondary than for its primary purpose. Rather, also in the light of Article 5(1)(b) GDPR, Article 6(4) GDPR specifies *additional* requirements for data re-use.[95]

This implies that Article 6(1) GDPR must be fulfilled, too, in practice mostly Article 6(1)(f) GDPR,[96] as well as potentially Article 9 GDPR. Again, there is significant uncertainty as to the application of this framework to data re-use for AI training, an aspect rightly highlighted by the Commission in its data strategy.[97] The last part of the paper will therefore develop guidelines which may serve as an interpretive framework for the GDPR, or as a blueprint for a new EU legal instrument on training data (below, V.2.b)i.).

---

[92] Schuhmacher and Fatalin, op. cit. *supra* note 65, pp. 203 et seq.
[93] European Commission, A European Data Strategy, COM(2020) 66 final, 6.
[94] See, e.g., Herbst, in: Kühling/Buchner (eds.), DS-GVO BDSG, 2nd ed. 2018, Art. 5 DS-GVO para. 28 et seq.; Buchner and Petri, in: Kühling/Buchner (eds.), DS-GVO/BDSG, 2nd ed. 2018, Art. 6 DS-GVO para. 183; Culik and Döpke, 'Zweckbindungsgrundsatz gegen unkontrollierten Einsatz von Big Data-Anwendungen', *Zeitschrift für Datenschutz* (2017), 226 (230).
[95] See also Article 29 Data Protection Working Party, 'Opinion 03/2013 on purpose limitation', WP 203, 2013, 12 n. 28.
[96] See Ursic and Custers, 'Legal Barriers and Enablers to Big Data Reuse' 2 *Eur. Data Prot. L. Rev*. (2016), 209 (212).
[97] European Commission, A European Data Strategy, COM(2020) 66 final, 6, 13, 17, 28 et seq.

b) Intellectual property law

Finally, it is conceivable that prospective training data is protected by copyright or related rights (e.g., the *sui generis* database right).[98] Such third-party rights may exist, for example, when works of fine art are used to train AI models who themselves create new works of art,[99] or when translation models are calibrated on legally protected templates of literature.[100] The training typically includes activities relevant for copyright protection. For example, the individual data must be saved on a server and stored in the working memory, which implies a reproduction of the work in terms of copyright.[101] If the original data is pre-processed, an adaptation relevant under copyright law[102] will often take place.[103] Finally, access to databases may involve an extraction requiring permission.[104]

As a consequence, the training can only be carried out in conformity with intellectual property rights if either a license is obtained or a specific exception is provided for the respective intellectual property right. Such an exception has now been enacted in the fully harmonizing Art. 3 et seq. of Directive 2019/790 on copyright in the digital single market (CDSM Directive). Hence, the question arises to what extent the European legislator has succeeded in achieving an adequate balance between the exploitation interests of the rightholders and innovation interests, i.e. to what extent the risk of blockage has been properly addressed.

i. The research TDM exception: Art. 3 CDSM Directive

According to Art. 3(1) CDSM-Directive, Member States must establish an exception to the right of reproduction and extraction when the use for text and data mining is made by 'research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access'. Thus, if such an actor has legally acquired access to the data, all acts of reproduction, but also pre-processing (e.g. normalization, see recital 8 CDSM Directive) are allowed for the purpose of automated data analysis.[105] This is essential because such pre-processing of data is typically required for machine learning.[106]

However, Art. 2(1) CDSM Directive defines the term 'research organisation' such that the organization must not operate for profit, with full reinvestment of profits in research or with a mission in the public interest recognized by the State. According to recital 12(7) CDSM

[98] Ursic and Custers, op. cit. *supra* note 96, p. 217 et seq.
[99] Overview in Mazzone and Elgammal, 'Art, Creativity, and the Potential of Artificial Intelligence', 8 *Arts* (2019), Article 26.
[100] See in detail, on the procedure, Rosati, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects, Briefing for the JURI committee of the European Parliament, 2018, 3 et seqq.
[101] Geiger, Frosio and Bulayenko, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Briefing for the JURI committee of the European Parliament, 2018, 6; Raue, 'Rechtssicherheit für datengestützte Forschung', *ZUM* (2019), 684 (685); Obergfell, 'Big Data und Urheberrecht', in: Ahrens et al. (eds.), *Festschrift für Wolfgang Büscher*, 2018, 223 (226); Spindler, 'Text und Data Mining', *GRUR* 2016, 1112 (1113); cf. also recital 8(6) and recital 9(2) CDSM Directive.
[102] See, e.g., Sec. 21 UK Copyright, Designs and Patents Act 1988; § 23 UrhG (German Copyright Act).
[103] Geiger, Frosio and Bulayenko, op. cit. *supra* note 101, p. 7; but see Obergfell, op. cit. *supra* note 101, p. 223 (226).
[104] BT-Drucks. 18/12329, 40; more nuanced Obergfell, op. cit. *supra* note 101, p. 227; database-specific questions can essentially be answered analogously to those genuinely related to copyright; see, e.g., Geiger, Frosio and Bulayenko, op. cit. *supra* note 101, p. 7; Raue, op. cit. *supra* note 101, p. 685.
[105] Raue, op. cit. *supra* note 101, pp. 687 et seq.; see also Spindler, 'Die neue Urheberrechts-Richtlinie der EU', *Computer und Recht* (2019), 277 (279).
[106] Kotsiantis/Kanellopoulos/Pintelas, op.cit. *supra* note 18, at 111.

Directive, organizations which are under the decisive influence of commercial enterprises are not covered. Therefore, profit-oriented companies, which are primarily examined in this article, cannot, even if they pursue research objectives and publish their results (as is not uncommon) in leading international journals,[107] invoke the implementation of Art. 3(1) CDSM Directive.[108]

ii. The general TDM exception: Art. 4 CDSM Directive

Commercial research and other profit-oriented uses of AI training are therefore only covered by Article 4 CDSM Directive. According to its first paragraph, a general limitation or exception 'for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining' must be established by Member States. As with Art. 3 CDSM Directive, there is no statutory right to remuneration for the rightholders (cf. recital 17 CDSM Directive).[109] However, according to Art. 4(3) CDSM Directive, the respective rightholders may exclude the application of this general TDM exception by expressly and appropriately declaring a reservation of the use of their protected works for TDM. In the case of content published online, this can be done, according to the wording of the provision, by means of machine-readable formats, for example. Given this veto right of rightholders, blocking possibilities continue to exist, as desired by the legislator.

IV. Assessment of the existing requirements: Coverage of the three risks in positive law

How, then, should the existing legal requirements be assessed with a view to adequately addressing the three risks of quality, discrimination and innovation?

1. Quality risks

Incentives to eliminate quality risks appear insufficient. While the above-mentioned data protection regulations could be used to create the basic framework of a quality regime for training data, it seems mostly likely that state-of-the-art anonymization strategies lead to the inapplicability of data protection law. Similar application issues as well as additional difficulties arise, as seen, with regard to product liability law.[110] Concerning contract law, it also seems more than questionable whether it can have a sufficient disciplinary effect. In particular, quality risks and deficiencies of AI applications, as will be explained in more detail below (V.1.a)i.), are typically difficult to recognize for purchasers,[111] and claims may quickly become time-barred according to the transpositions of Article 5 of the Consumer Sales Directive.[112] The steering effect of existing law is therefore seriously limited when it comes to quality risks in training data.

2. Discrimination risks

The risk of discrimination is not yet properly addressed in terms of training data, either. It is true that anti-discrimination law prohibits unjustified discrimination even on the basis of AI models. However, pure command-and-control regulation of AI results is likely to be inadequate,

[107] See only the references *supra* note 12.
[108] More precisely Raue, op. cit. *supra* note 101, p. 690.
[109] Spindler, op. cit. *supra* note 105, p. 281.
[110] See also Raue, op. cit. *supra* note 65, p. 1845.
[111] Butterworth, 'The ICO and artificial intelligence: The role of fairness in the GDPR framework', 34 *Computer Law & Security Review* (2018), 257, 261.
[112] Cf. on the latter Raue, op. cit. *supra* note 65, p. 1843.

a point also emphasized by the EU Commission.[113] On the one hand, enforcement deficits and problems of proof lead, as seen, to a considerable loss of incentives; on the other hand, victims of discrimination law only have ex post corrective instruments, such as claims for damages, at their disposal. This implies, however, that damage has already been suffered by the victims, which can be quite significant (even in immaterial dimensions), particularly in the area of discrimination.

As a result, anti-discrimination law tends to come too late. Although data protection law could provide ex ante mechanisms, for example through audits in accordance with Article 58(1)(b) GDPR, its applicability at the training data stage is dubious in view of widespread anonymization. Hence, a regulatory regime for training data should abstract away from the always controversial question of identifiability in terms of data protection law (see below, V.2.a)).

3. Innovation risks

Only the innovation risk, in the form of a blocking risk stemming from intellectual property rights, has recently found a concrete solution in the TDM exceptions. Art. 4(1) CDSM Directive establishes a default rule in favor of the (commercial) use of protected data for training purposes, with a simultaneous opt-out option for the rightholders in Art. 4(3). This reverses the burden of activity: whereas users of protected content normally have to approach the rightholders to conclude a license agreement, the rightholders themselves now have to take action. Behavioural economic effects such as the status quo bias[114] suggest that the opt-out mechanism will lead to a significantly higher number of works that can be used for training purposes than an, alternatively conceivable, opt-in mechanism. At first glance, the reversal of the burden of activity appears to be an appropriate balance between exploitation and innovation interests.

However, it should be borne in mind that a TDM opt-out can be declared with little effort, so that the effect of the status quo bias remains to be seen. In particular, commercial research, which is currently particularly important in the context of AI, but which, according to the wording of the CDSM Directive, cannot invoke the more generous Article 3 exception, therefore stands to benefit from the regulation only to a limited extent. This gives rise to a certain need for adaptation (see below, V.2.b)ii.). Overall, however, it must be positively acknowledged that the legislator has now addressed a specific risk of AI training, copyright blockage, with significantly greater precision than in data protection, anti-discrimination or liability law.

The downside of this acknowledgment is that, while the GDPR does contain a framework for the re-use of data covered by data protection law, it operates with highly vague balancing tests. These will need to be specified in order to give AI developers the interpretive tools and the incentives necessary for legally compliant innovation (see below, V.2.b)i.).

V. Prospects for reform: Toward a comprehensive legal framework for training data

This assessment provides guidance for a reform agenda, which is spelled out in this final part of the paper. However, in order to specify legislative measures with regard to the three risks

[113] European Commission, op. cit. *supra* note 1, pp. 23-25.
[114] Samuelson and Zeckhauser, 'Status quo bias in decision making', 1 *Journal of Risk and Uncertainty* (1988), 7.

mentioned, regulatory foundations must first be laid, in all due brevity (1.). Concrete proposals for addressing the three risks can then be examined (2.).

1. Regulatory foundations

Regulation is not appropriate if the risks examined are adequately addressed by market solutions or if the costs of regulation would be too high.[115] However, a brief economic analysis suggests that neither sufficient market solutions nor prohibitive regulatory costs exist.

a) Market failure

First, it is not apparent that the regulatory risks mentioned above could be completely resolved by pure market solutions. This is generally supported by the fact that AI models have already been trained with large data sets for several years, at least since the breakthroughs in the area of deep learning in 2006,[116] without the risks having lost any of their topicality.

i. Quality risks

In the area of quality risks, information asymmetries in particular stand in the way of a market solution. This is because the quality of an AI model can typically only be accurately estimated by the developers. There is a whole range of performance metrics (*accuracy*, *precision*, *recall*, *F1 score,* etc.)[117] that indicate the predictive quality achieved by each model. However, these metrics are not always published in commercial applications. Furthermore, these measures usually only refer to the so-called *test performance.*[118] Training data is split into two data sets for this purpose: The model is trained using one set of data, and the performance is then tested on the held-out data.[119] However, this means that the performance metrics only indicate how well the model operates on the test data set. Depending on the representativeness of this data set for the actual conditions of use, there can be considerable deviations between *test* and *field performance.*[120] The degree to which a model generalizes from the test data set to field use is typically best assessed by the developers.[121] However, they have little economic interest in disclosing any quality risks.

In addition, field performance is often difficult to measure because it is only possible to determine whether the model has made an error or not for a small proportion of the cases actually examined by the model (the positively selected cases, so-called *reject inference*).[122] If, for example, only one of the 500 candidates ranked by a recruitment tool is hired, it is impossible to say with hindsight whether one or more of the remaining 499 candidates might have performed better in the job advertised.[123] AI applications can therefore represent *credence*

---

[115] See only Veljanovski, 'Economic Approaches to Regulation', in: Baldwin/Cave/Lodge (Eds.), *The Oxford Handbook of Regulation*, 2010, 18 (20 et seqq.).
[116] Fundamentally Hinton, Osindero and The, 'A fast learning algorithm for deep belief nets', 18 *Neural Computation* (2006), 1527; overview in Goodfellow, Bengio and Courville, op. cit. *supra* note 4, p. 18.
[117] Goodfellow, Bengio and Courville, op. cit. *supra* note 4, p. 100 et seq., 410 et seqq.; particularly on credit scoring Hand, 'Good practice in retail credit scorecard assessment', 56 *Journal of the Operational Research Society* (2005), 1109 (1111 et seqq.).
[118] Hand, op. cit. *supra* note 28, p. 2 et seq.
[119] LeCun, Bengio and Hinton, op. cit. *supra* note 9, p. 437.
[120] Goodfellow, Bengio and Courville, op. cit. *supra* note 4, p. 107; Hand, op. cit. *supra* note 28, p. 7 et seq.
[121] Cf. Hand, op. cit. *supra* note 28, p. 9.
[122] Hand, op. cit. *supra* note 28, pp. 3, 9; Hand, op. cit. *supra* note 117, p. 1116.
[123] Kim, 'Data-Driven Discrimination at Work', 58 *William & Mary Law Review* (2017), 857 (894 et seq.); Hacker, op. cit. *supra* note 23, p. 1150.

goods[124] for which a regulatory quality assurance regime that complements the market also makes sense from a law-and-economics perspective.[125]

ii. Discrimination risks

Discrimination risks also tend to be inadequately addressed by market forces.[126] This is not only evidenced by the fact that discrimination can be statistically efficient.[127] Furthermore, it also produces undesirable economic feedback effects: biased training data tend to further increase the marginalization of already disadvantaged groups and the prioritization of already preferred groups by what may be termed a machine-mediated self-fulfilling prophecy.[128] These feedback effects seriously question the efficiency and inclusiveness of AI-based analysis systems.

iii. Innovation risks

Finally, the innovation risks associated with the possibility of being blocked by existing intellectual property and data protection rights cannot be solved efficiently by the market, either. In view of the large amount of data and the large number of rightholders involved, negotiated solutions fail because of prohibitive transaction costs.[129] From a law-and-economics perspective, this is precisely the reason for the establishment of copyright exceptions.[130]

b) Costs of regulation

Even if pure market solutions fail, a law-and-economics perspective suggests that regulation should only be enacted if the expected costs of regulation are lower than the expected benefit.[131] However, these costs and benefits are extremely difficult to quantify, especially in the case of intangible harms to data protection and non-discrimination.[132] Therefore, in the case of training data, it seems necessary to merely require that the expected regulatory costs be proportionate to the risks addressed. In this way, the risk of over-regulation that could endanger innovation can be contained. Ultimately, however, this requirement must be substantiated in each of the individual measures proposed in the following section.

2. A regulatory framework for specific risks

On this basis, the last part of the paper develops a regulatory framework for training data. The focus is first on quality and discrimination risks (a). With regard to the innovation risks resulting from possible blocking (b), only modest modifications are suggested concerning the copyright regime of the CDSM Directive, but more detailed guidelines are advanced concerning the GDPR.

---

[124] Fundamentally on credence goods Darby and Karni, 'Free Competition and the Optimal Amount of Fraud', 16 *Journal of Law and Economics* (1973), 67; on computer specialists as providers of credence goods Dulleck and Kerschbamer, 'On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods', 44 *Journal of Economic Literature* (2006), 5.

[125] More precisely Dulleck and Kerschbamer, op. cit. *supra* note 124, p. 15 et seqq.

[126] See Pasquale, op. cit. *supra* note 13, p. 1926.

[127] Romei and Ruggieri, 'A multidisciplinary survey on discrimination analysis', 29 *The Knowledge Engineering Review* (2014), 582 (592 et seq.); more in detail, and nuanced, Schwab, 'Is statistical discrimination efficient?', 76 *The American Economic Review* (1986), 228.

[128] Kim, op. cit. *supra* note 123, p. 895 et seq.; Hacker, op. cit. *supra* note 23, p. 1150.

[129] Cf. Ursic and Custers, op. cit. *supra* note 96, p. 213.

[130] Gordon, 'Fair Use as Market Failure' 82 *Colum. L. Rev.* (1982), 1600 (1613 et seqq.).

[131] Veljanovski, op. cit. *supra* note 115, p. 22.

[132] Cf. Keat, 'Values and Preferences in Neo-Classical Environmental Economics', in: Foster (ed.), *Valuing Nature?*, 1997, 32 (39-42); Mishan and Quah, Cost-Benefit Analysis, 5th ed., 2007, 179 et seqq.

a) Quality and discrimination risks

From a policy perspective, the starting point for the treatment of quality and discrimination risks is that these two risks are often so closely interwoven that they should not be considered separately and subjected to disparate regulations, but should be treated by a single piece of regulation concerning training data. Only in this way can a coherent, discrimination-sensitive quality assurance law of algorithmic processes be created.[133] At the same time, such a regime must, as seen, become independent of the question of identifiability and develop overarching criteria for training data and environments.

i. Data quality and data balance

In order to address quality and discrimination risks, however, it must first be clarified what constitutes data quality in the area of training data and how discrimination can technically arise from them. In recent years, the computer science literature has developed a whole catalogue of criteria and metrics for data quality[134] and the discrimination potential[135] in data sets, and even the ISO 25012 standard for data quality.[136]

(1) Accuracy

Article 5(1)(d) GDPR provides a first indication of such a quality regime, even if, as seen, it is not necessarily applicable to training data. However, it shows that the accuracy of data is an important dimension of data quality. This is quite undisputed in computer science research.[137] Factually incorrect training data are crucial, if only because they render the result of an AI model wrong even if the actual input data of the person being analyzed is correct.[138] However, concrete accuracy metrics will have to be agreed upon. As discussed, they will need to take into consideration the size of the data set and define acceptable margins of error in relation to the context in which the trained model is deployed.

(2) Timeliness

A second element, also contained in Article 5(1)(d) GDPR, is the timeliness of the data. This criterion is also agreed upon in the computer science literature.[139] Preferences, contexts and social norms change over time. However, the relevance of these changes differs between data

---

[133] Cf. also Gerberding and Wagner, op. cit. *supra* note 15, p. 117 with the demand for the development of a quality assurance law for scoring algorithms.

[134] Overview by Lee et al., 'AIMQ: a methodology for information quality assessment', 40 *Information & Management* (2002), 133 (134 et seqq.); Heinrich and Klier, 'Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement', in: Hildebrand et al. (Ed.), Daten- und Informationsqualität, 4th ed., 2018, 47 (50 et seqq.), each with an emphasis on completeness, accuracy, consistency and timeliness; fundamentally Wang, Storey and Firth, 'A Framework for Analysis of Data Quality Research', 7 *IEEE Transaction on Knowledge and Data Engineering* (1995), 623.

[135] See, e.g., Calders and Žliobaitė, op. cit. *supra* note 28; Romei and Ruggieri, op. cit. *supra* note 127, p. 582; Žliobaitė, 'Measuring discrimination in algorithmic decision making', 31 *Data Mining and Knowledge Discovery* (2017), 1060.

[136] ISO/IEC 25012, https://iso25000.com/index.php/en/iso-25000-standards/iso-25012, with the dimensions of accuracy, completeness, consistency, credibility and currentness.

[137] See only Heinrich and Klier, op. cit. *supra* note 134, pp. 55-57; Lee et al., op. cit. *supra* note 134, p. 134; Wang, Storey and Firth, op. cit. *supra* note 134, pp. 628 et seq.; Hoeren, op. cit. *supra* note 19, pp. 10 et seq.

[138] See also the Sachverständigenrat für Verbraucherfragen, op. cit. *supra* note 19, p. 46 n. 34.

[139] Heinrich and Klier, op. cit. *supra* note 134, p. 59 et seq.; Lee et al., op. cit. *supra* note 134, p. 134; Hand and Henley, 'Statistical Classification Methods in Consumer Credit Scoring: a Review', 160 *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (1997), 523 (525); Wang, Storey and Firth, op. cit. *supra* note 134, p. 628 et seq.

types.[140] For some data, timeliness is of considerable importance.[141] It is well-known, for example, that the use of historical data sets can contribute to the perpetuation of forms of structural discrimination that have been overcome in the meantime but were more pronounced in the past (*historical bias*).[142] In this respect, yesterday's data must not drive tomorrow's decisions. On the other hand, there are data types for which even older data lose little or no significance;[143] in this respect, one only has to think of medical test series.

(3) Completeness and factor diversity

Third, completeness and factor diversity in training data are desirable. AI training data for supervised learning consists of (numerical or categorical) values assigned to a number of decision factors (so-called *features*, e.g. age, shoe size, income). The number of these features ranges from one to many, and they are all weighted differently.[144] First of all, values should be available for all features and all individuals, i.e. all entry possibilities should be filled in a training data set (completeness).[145]

Furthermore, increasing the diversity of factors (in the sense of the number of factors that are not closely correlated with each other[146]) may reduce the likelihood of the output being closely correlated with membership in protected groups.[147] This suggests that the number of independent factors in training data should exceed a lower threshold (for example, five). However, since even such factor diversity cannot, in all cases, prevent close correlations of the output with group membership,[148] it seems reasonable to implement this requirement merely as a non-binding target rule. In any case, AI developers would be generally free to weight the factors, so that the intensity of intervention would be limited.

It remains questionable then to what extent, in the case of training data which – contrary to the target rule – is based on only one or very few factors, those factors should be excluded which are known to closely correlate with membership in protected groups. Such a rule can help to prevent statistical discrimination.[149] This insight is the reason behind § 31(1) no. 3 BDSG,

[140] Information Commissioner's Office, 'Principle (d): Accuracy', https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/accuracy/; Heinrich and Klier, op. cit. *supra* note 134, p. 60.
[141] See, for example, the references *supra* note 139 on credit scoring; more generally Hand, op. cit. *supra* note 28, p. 7 et seq.
[142] See Calders and Žliobaitė, op. cit. *supra* note 28, p. 48 et seq.; Hacker, op. cit. *supra* note 23, p. 1148; and the references *supra* note 25.
[143] Calders and Žliobaitė, op. cit. *supra* note 28, p. 48.
[144] Goodfellow, Bengio and Courville, op. cit. *supra* note 4, p. 96.
[145] Heinrich and Klier, op. cit. *supra* note 134, p. 52; Lee et al., op. cit. *supra* note 134, p. 134; Wang, Storey and Firth, op. cit. *supra* note 134, p. 628.
[146] On formal diversity concepts, see Drosou and Pitoura, 'Multiple Radii DisC Diversity', 40 *ACM Transactions on Database Systems (TODS)* (2015), Article 4, 1 (1 et seq.); on the importance of factor analysis Hair et al., *Multivariate Data Analysis*, 9th ed. 2019, 121 et seqq.
[147] Schröder et al., 'Ökonomische Bedeutung und Funktion Funktion von Credit Scoring', in: Schröder/Taeger (eds.), *Scoring im Fokus*, 2014, 8 (42); but see on problems with high-dimensional feature spaces (multi-collinearity) Hair et al., op. cit. *supra* note 146, p. 311 et seqq.
[148] Just think of five features used, which are independent of each other, but all correlate closely with group membership.
[149] In detail Britz, *Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung* [Case-By-Case Justice versus Generalization. Constitutional Limits of Statistical Discrimination], 2008, 120 et seqq.

mentioned above, according to which scoring must not be based exclusively on address data.[150] On the other hand, it must be taken into account that unifactorial modelling, when there is a very close correlation between factor and group membership, has the effect of direct discrimination.[151] However, even direct discrimination can be justified under certain circumstances in EU law.[152] Therefore, mandatory prohibitions for training data with few factors should ultimately be rejected, and possible risks of discrimination be solved via existing anti-discrimination law.

(4) Balance

A fourth quality criterion, which is particularly relevant for the prevention of discrimination, but which has not yet been adequately covered by regulation, is the balance of the data set between different groups protected under anti-discrimination law. The Commission White Paper also lists this criterion.[153] Here, too, empirical studies have shown that an imbalance caused by over- or under-representation of individual groups can lead to a deterioration in the prediction quality for protected groups and ultimately to systematically negative distortion (*sampling bias*).[154] From a technical perspective, certain possibilities for re-balancing data sets do exist.[155]

(5) Representativeness

Finally, data quality also includes the representativeness of the data for the target context,[156] as underlined both by the Commission's White Paper and the accompanying Liability Report.[157] Representativeness overlaps with the criterion just mentioned in so far as a lack of balance can, depending on the target context, but need not lead to a lack of representativeness. Moreover, the latter term is broader since it is not limited to the attributes protected by anti-discrimination law – just think of socio-economic differences.[158]

The relevance of representativeness also extends beyond supervised learning: attention must be paid to representativeness of the training environment in reinforcement learning settings, too. It is an issue concerning both the quality of the model and non-discrimination if certain groups are severely underrepresented in the learning environment. An extreme example (also regarding the balance of protected groups) would go as follows: if the control AI of an autonomous vehicle is mainly confronted with people of white skin color during training, it may not recognize people of darker skin color, or recognizes them less often, as people. An empirical study with precisely these results shows that this concern is not unfounded.[159]

---

[150] See also, on bias potentially introduced by reliance on postal codes, Kroll et al. 'Accountable Algorithms', 165 *U. Pa. L. Rev.* (2016), 633 (681, 685); Kamarinou, Millard and Singh, 'Machine learning with personal data', Queen Mary School of Law Legal Studies Research Paper 247/2016, 16.
[151] See Thüsing, in: Münchener Kommentar, AGG, 8th edition 2018, § 3 para. 15.
[152] See, for a discussion, Ellis and Watson, op. cit. *supra* note 87, 171-174, 381 et seqq.
[153] European Commission, op. cit. *supra* note 1, p. 19.
[154] See the references *supra* note 28 and the cases *supra* note 24.
[155] See, e.g., Zemel et al., 'Learning Fair Representations', *Proceedings of the 30th International Conference on Machine Learning* (2013), 325; Wang et al., 'Balanced Datasets Are Not Enough', *Proceedings of the IEEE International Conference on Computer Vision* (2019), 5310.
[156] Hand, op. cit. *supra* note 28, p. 8 et seq.; Sachverständigenrat für Verbraucherfragen, op. cit. *supra* note 19, p. 145.
[157] European Commission, op. cit. *supra* note 1, p. 19; European Commission, op. cit. *supra* note 3, p. 8.
[158] See for instance Pasquale, op. cit. *supra* note 13, pp. 1923 et seq.
[159] Wilson, Hoffman and Morgenstern, 'Predictive inequity in object detection', Working Paper, 2019, https://arxiv.org/abs/1902.11097.

ii. Regulatory implementation

The five quality criteria mentioned above establish an ideal vision that can hardly be fully achieved under real conditions. This must be taken into account in any regulatory implementation. For example, it would probably be prohibitively expensive to create a training data set with annotated facial images containing exactly the same number of individuals from all ethnic groups in the world. Although there are currently efforts in the industry in this direction,[160] they must necessarily remain approximate.

(1) Possible measures

Concrete implementation measures to ensure data quality and non-discrimination may, for example,[161] consist of continuous monitoring and testing,[162] such as random sampling to detect objectively incorrect data,[163] the use of error[164] and bias correction algorithms,[165] iterative updates of and, where appropriate, 'expiry dates' for certain data sets that are subject to particularly strong temporal changes. Other useful procedural requirements include the documentation[166] and publication of the provenance of training data (to determine distortions caused by historical data),[167] of meta-data with regard to the training data set (e.g. descriptive statistics)[168] and of state-of-the-art, possibly standardized performance metrics.[169] Such mandatory disclosure would also counteract the aforementioned information asymmetry between developers and buyers.

(2) A risk-based approach

From a legal point of view, the crucial question is, therefore, to what extent these measures should be prescribed by regulation. Here, the costs or the effort for the implementation of the individual measures will have to be put in relation to the risks associated with the respective application.[170] The Commission's White Paper also proposes such a risk-based approach,[171] as does the report of the German Data Ethics Commission.[172]

---

[160] Yang et al., 'Towards Fairer Datasets', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)* (2020), 547; Google Research, 'Inclusive Images Challenge', kaggle (2018), https://www.kaggle.com/c/inclusive-images-challenge.

[161] See the overview in Pasquale, op. cit. *supra* note 13, p. 1932 et seqq.

[162] ACM, 'Statement on Algorithmic Transparency and Accountability and Principles for Algorithmic Transparency and Accountability, 2017, https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf, Principle 7; Schröder et al., op. cit. *supra* note 147, p. 45.

[163] Cf. Diakopoulos et al., 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms', Working Paper, 2018, Fairness, Accountability, and Transparency in Machine Learning, https://www.fatml.org/resources/principles-for-accountable-algorithms, under "Accuracy"; Sachverständigenrat für Verbraucherfragen, op. cit. *supra* note 19, p. 83.

[164] See, e.g., Schröder et al, op. cit. *supra* note 147, p. 28.

[165] See, e.g., Zehlike, Hacker and Wiedemann, 'Matching code and law: achieving algorithmic fairness with optimal transport', 34 *Data Mining and Knowledge Discovery* (2020), 163 and the overview *supra* note 31.

[166] For further documentation requirements in the ML pipeline, see Selbst and Barocas, 'The Intuitive Appeal of Explainable Machines', 87 *Fordham Law Review* 2018, 1085 (1130 et seqq.); Hacker, 'AI regulation at the European and the National Level', Working Paper, 2020.

[167] ACM, op. cit. *supra* note 162, Principle 5.

[168] European Commission, op. cit. *supra* note 1, p. 19; Selbst and Barocas, op. cit. *supra* note 166.

[169] Cf. Deussen et al., op. cit. *supra* note 21, p. 21.

[170] Similarly Deussen et al., op. cit. *supra* note 21, p. 6.

[171] European Commission, op. cit. *supra* note 1, p. 17.

[172] German Data Ethics Commission, op. cit. *supra* note 20, p. 173 et seq.

In order to determine the concrete strictness of the regulatory requirements, first, sector-specific (vertical) distinctions should be made; the areas to which EU anti-discrimination law applies can provide an indication of particularly risky sectors.[173] In addition, any existing market solutions that could speak in favor of lowering regulatory requirements for certain applications must be specifically evaluated. The Commission also rightly emphasizes that, even in high-risk sectors, the nature of the (intended) concrete use of the AI model must be taken into account as well.[174] Second, it is worth considering the possibility of additionally and sector-independently (horizontally) covering particularly risky forms of AI applications, with strict prerequisites. An example, also mentioned in the White Paper,[175] is face recognition software (*remote biometric identification*).

Overall, the legal framework for training data should therefore be committed to the process of risk-based regulation now also taken up in the GDPR.[176] At the same time, this constitutes a core problem: different sectors and applications need to be assigned to different risk levels. Three examples: Autonomous driving should be placed in the highest category because of the associated dangers to life and limb;[177] AI recruitment in an intermediate class because of the considerable impact on income and lifestyle of candidates;[178] and personalized advertising in a low category because of the relatively limited disadvantages resulting from incorrectly targeted advertising. However, there is still a considerable need for research with regard to the exact risk classification of different applications.[179] Ultimately, it will not always be necessary to spell out a classification explicitly; it may be more effective, and more conducive to legislative consensus, to differentiate implicitly via sectoral and application-related requirements without necessarily allocating each sector or application to a specific, and rather abstract, risk class.

iii. Claims of affected persons

A final aspect of the analysis with respect to quality and discrimination risks is the linking of regulatory requirements with possible claims by affected persons. In addition to public law enforcement of the regulatory framework, decentralized private enforcement should not be neglected.[180] Here the focus is on liability on the one hand (1) and on access rights on the other (2).

(1) Liability

As far as liability is concerned, the mentioned regulatory requirements should also function as minimum standards for safety obligations under tort law.[181] In addition, a violation of the measures required by the regulatory provisions should trigger a rebuttable presumption, also in civil proceedings, that (i) a discriminatory result is causally attributable to biased training data and (ii) that a defect within the meaning product liability was negligently caused by inadequate training data. This is crucial in so far as causal processes are generally difficult to illuminate in

[173] See also the examples in German Data Ethics Commission, op. cit. *supra* note 20, pp. 177 et seqq.
[174] See European Commission, op. cit. *supra* note 1, p. 17 and the examples there.
[175] European Commission, op. cit. *supra* note 1, p. 18, 21 et seq.
[176] See the references *supra* note 57.
[177] In this sense also European Commission, op. cit. *supra* note 1, p. 17.
[178] But see European Commission, op. cit. *supra* note 1, p. 18 (high risk).
[179] More specifically, Hacker, op. cit. *supra* note 166.
[180] Pasquale, op. cit. *supra* note 13, p. 1920; Lyndon, 12 *Yale J. on Reg.* (1995), 137 (143).
[181] Wagner, in: Münchener Kommentar, BGB, 7th ed. 2017, § 823 para. 447 et seq.; see also Zech, op. cit. *supra* note 73, p. 211, for IT security requirements.

the AI arena.[182] In the case of discrimination, biased training data significantly speaks against justification.[183] In the area of product liability, the presumption of causality facilitates redress against AI developers (who may be different from the manufacturer).[184] Importantly, again, the scope of product liability, to effectively cover AI, must be extended to software.[185]

These presumptions, in turn, increase the incentives to comply with the proposed regulatory requirements for training data. Such an extension of the violation of regulatory requirements to claims for damages is not a legal novelty, either. It is inherent, for example, in Article 9(1) of the Market Abuse Regulation[186] and has been suggested, for different types of regulation, by the Expert Group on Liability and New Technologies as well.[187]

(2) Access rights

With regard to the rights of potential victims to access information, it must be considered whether the restrictive line of the CJEU case law in *Meister*, mentioned above, should be corrected and whether, therefore, it ought to be possible for affected parties to request information on certain parameters of the training data in the case of a justified initial suspicion of quality defects or discrimination. This makes sense especially for aggregated information on score distributions between protected groups, to prove statistical inequality of treatment. Affected parties should be granted an access right in this case. Arguably, however, Article 15 of the GDPR may close some gaps in the case of personal data.[188]

Overall, the information interests of potentially injured parties must be weighed not only against the right to data protection of potentially identifiable third parties, but also against the legitimate confidentiality interests of AI developers, in order to prevent unreasonable innovation risks. Thus, any rights to information must also be coordinated with the Trade Secrets Directive,[189] in particular the catalogue of exceptions in its Article 5, which can serve as a model for the above-mentioned balancing exercise.

b) Innovation risks

Turning finally to innovation risks, the Commission is entirely right in flagging the problem of the re-use of data as one of the main challenges for a European data strategy, and for innovation based on AI in general.[190] Besides the problem of access to data, which has been addressed for public sector data by the Open Data Directive[191] and cannot be developed in detail here,[192] the potential protection of the training data by data protection and intellectual property rights proves to be the main legal obstacle to data re-use.

---

[182] Expert Group on Liability and New Technologies, op. cit. *supra* note 88, 20 et seq.; Zech, op. cit. *supra* note 83, p. 52 et seq.; Zech, op. cit. *supra* note 73, p. 205 et seqq., in particular 208, 217.
[183] Hacker, op. cit. *supra* note 23, p. 1163 et seq.; Schönberger, op. cit. *supra* note 74, p. 184 et seq.
[184] See European Commission, op. cit. *supra* note 3, p. 14.
[185] See the references *supra* note 80 et seq.
[186] See, e.g., Grundmann, in: Staub, HGB, 5th edition 2016, vol. 11/1, Bankvertragsrecht 6th Part, para. 401.
[187] Expert Group on Liability and New Technologies, op. cit. *supra* note 88, 47 para. 22, 48 para. 24.
[188] Hacker, op. cit. *supra* note 23, p. 1173 et seq.
[189] Directive (EU) 2016/943.
[190] European Commission, A European Data Strategy, COM(2020) 66 final, 6, 13, 17, 28 et seq.
[191] Directive (EU) 2019/1024.
[192] See only Rubinstein and Gal, 'Access Barriers to Big Data', 59 *Ariz. L. Rev*. (2017), 339; Ursic and Custers, op. cit. *supra* note 96, p. 215 et seq., 218 et seqq.

i. Towards a clarified data protection regime

Concerning data protection law, I have argued above that strong anonymization techniques should put the data set beyond the scope of the GDPR. However, until a clarifying ruling by the CJEU, legal insecurity concerning this question will lead many AI developers, as a precautionary compliance measure, to assume the applicability of the GDPR to their training data set. Therefore, any new legal instrument covering training data must address the question of re-use under data protection law. In the meantime, under the GDPR, guidelines should be developed by the European Data Protection Board tackling this question (Art. 70(1)(e) GDPR).

Unfortunately, there is neither a silver bullet available nor can a bright red line be drawn which would distinguish legal from illegal re-use for training data purposes under the GDPR. While much will depend on the concrete circumstances, the following general criteria should be decisive for an analysis under Articles 6(1)(f), 6(4) and 9 GDPR.

(1) Guidelines for Article 6(1)(f) GDPR

Since transaction costs for securing consent of each data subject represented in the training data set will often be prohibitive,[193] the key legal basis for training an AI model with personal data will be Article 6(1)(f) GDPR. Here, one must strictly distinguish between the training operation on the training data set and the consecutive analysis of new data subjects with the help of the trained model.

As regards the training itself, the interests of the controller and of third parties have to be weighed against those of the data subjects represented in the data set. Clearly, important factors will be the degree of anonymization,[194] the wider social benefits expected from the model, and the proximity of the data used to sensitive categories of Article 9 GDPR.[195] The decisive element, in my view, however, should be the extent to which the training operation itself adds new data protection risks for the data subjects. In fact, it is submitted that in a supervised learning strategy, these risks are typically very small. This is because the training of the model does not reveal any new information concerning the data subjects contained in the training data: it is precisely because the target qualities are already known that supervised learning can be conducted in the first place. For example, let us imagine that a lender has a data set concerning three categories: default events; degree of education; and yearly income. Using the latter two features, the lender wants to build a model predicting the risk of default events, i.e., a credit score. In supervised learning, it will use the information about known default events of the data subjects in the training data to calibrate (supervise) the model.[196] While the model will discover potentially novel relationships between the feature variables (education, income) and the default risk, the training operation itself does not reveal anything substantially new about the default risk of the subjects in the data set. Rather, their default events are treated as 'ground truth' to correct the model.[197]

Therefore, the only real risk for the data subjects represented in the training data consists in IT security risks that may be increased if, for the purposes of training, the data set is copied or

[193] Mészáros and Ho, 'Big Data and Scientific Research', 59 Hungarian Journal of Legal Studies (2018), 403 (405); Ursic and Custers, op. cit. *supra* note 96, p. 213.
[194] Hintze, op. cit. *supra* note 42, at 94 et seq.
[195] Article 29 Working Party, 'Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC', 2014, WP 217, 37 et seq.
[196] See the references *supra* note 9.
[197] See Shalev-Shwartz and Ben-David, op. cit. *supra* note 4, p. 4.

moved to new storage locations. These risks must be properly addressed, particularly through Articles 32 et seqq. GDPR; but they will not generally be so important as to flatly outweigh the interests of the model developer and of third parties. In this sense, training the model, from a data protection perspective, is similar to data mining from a copyright perspective: it can be equated to reading the data anew, without generating substantial new risks for those present in the data set. Hence, in data protection law, too, the motto should be: 'the right to train is the right to read'.[198] Contrary to the existing literature,[199] this suggests a rather permissive understanding of Article 6(1)(f) GDPR for data re-use.

This understanding, however, does not prejudice the entirely different question of the legality of *applying* the trained model to new data subjects, in the field, e.g. to assess their credit risks. Here, Article 22 GDPR, for example, may enter the game. Similarly, if the data set is used for unsupervised learning to discover *new* relationships between the data subjects (e.g., clustering),[200] new risks may arise from this novel information. Finally, the training data set may be acquired by the training entity before conducting the modeling, which again triggers new risks due to the data transmission.[201] These are separate questions which, while important for AI practice in general, transcend the scope of this paper.

(2) Guidelines for Article 6(4) GDPR

As seen, Article 6(4) GDPR contains specific and additional provisions for the secondary use of data. It specifies a compatibility test which must take into account the following criteria: (a) the link between the primary and the secondary use; (b) the data collection context; (c) the proximity of the data to sensitive categories; (d) the consequences of the secondary use for data subjects; and (e) the existence of safeguards, including encryption and pseudonymization.

Concerning data re-use for training, we have just seen that the consequences for data subjects, with respect to data protection risks, are typically quite limited. Therefore, if state-of-the-art pseudonymization or anonymization techniques are deployed, the training itself should pass muster under Article 6(4) GDPR. This should hold even if the link between the primary and the secondary use is weak: for the data subject, it should not matter if this link is strong or weak as long as the risks entailed are low. For research and statistics, this is explicitly provided for in Art. 5(1)(b) GDPR.[202]

(3) Guidelines for Article 9 GDPR

The risk-based approach just advanced should also determine the treatment of AI training under Article 9 GDPR. As is well-known, there is no general balancing test mirroring Article 6(1)(f) GDPR for sensitive data. However, given the relatively low risks involved with the training operation itself, developers should be able to avail themselves rather generously of the public interest clause contained in Article 9(2)(g) GDPR, for example if the model is consciously trained to foster to legal equality (Art. 20 of the Charter of Fundamental Rights) and non-

[198] On the copyright policy demand 'the right to mine is the right to read', see Murray-Rust, Molloy and Cabell, 'Open Content Mining', in: Moore (ed.), *Issues in Open Research Data*, 2014, 11, 27 et seq.; Geiger, Frosio and Bulayenko, op. cit. *supra* note 101, p. 21.
[199] For a more restrictive understanding, see, e.g., Ursic and Custers, op. cit. *supra* note 96, p. 212 et seq.
[200] See Goodfellow, Bengio and Courville, op. cit. *supra* note 4, p. 102.
[201] Information Commissioner's Office, 'Royal Free - Google DeepMind trial failed to comply with data protection law (July 3, 2017); Mészáros and Ho, op. cit. *supra* note 193, at 406.
[202] Cf. Kotschy, 'Lawfulness of processing', in: Kuner et al. (eds.), The EU General Data Protection Regulation (GDPR). A Commentary, OUP (forthcoming), https://works.bepress.com/christopher-kuner/1/download/, at 54.

discrimination (Art. 21 of the Charter),[203] e.g. by attempting to mitigate bias in hiring processes. Again, this result holds only for the training operation, not for the field application.

To the extent, however, that the AI model is built for research purposes (e.g., to predict cancer risk), Article 9(2)(j) and Article 89 GDPR provides Member States with leeway to develop particular, more tailored rules. While the UK legislator has provided details on medical research (Sec. 19 UK DPA 2018),[204] the German legislator, for example, has introduced a new § 27 BDSG which, in its first paragraph, contains a specific balancing test for sensitive data. Commentators agree that the rule is more restrictive for developers than Article 6(1)(f) GDPR,[205] as the interests of the controller must 'significantly outweigh' those of the data subject.[206] However, given the low risks arising from the training itself, even this threshold can arguably often be passed.

(4) Brief summary

In sum, the guidelines suggested here should take into account the relatively low risks involved with the (supervised) training process of an AI model itself. Under a risk-based approach, therefore, data re-use for training purposes should be treated more permissively under the GDPR than generally assumed. Importantly, Article 89 GDPR (and § 27 BDSG) must be read, in the light of recital 159 GDPR, to privilege both commercial and non-commercial research.[207] This directly links to the discussion of the TDM exception in copyright law, where this distinction plays a much greater role.

ii. Copyright and the TDM exception

Regarding the risks of innovation resulting from the possible blockage by intellectual property rights, an innovation-friendly interpretation of the TDM exception should be ensured. However, it has to consider that innovation interests reside on both sides of the aisle – that of the developers and that of the rightholders (cf. the second recital of the CDSM Directive). Nevertheless, situations could arise where AI developers are in urgent need of training their model on certain data, but the rightholders opportunistically demand substantial remuneration for withdrawing their veto under Article 4(3) CDSM Directive, despite the fact that their economic interests are only marginally affected (similar to the well-known hold-up problem in long-term contracting[208]).

Hence, the question arises, particularly with regard to the substantially different treatment of non-commercial and commercial research, whether, in view of the principle of equal treatment enshrined in Article 20 of the Charter, commercial research should not also benefit from Article 3 CDSM Directive, at least in certain constellations of technical or economic necessity for training on certain data sets. The Article 3 exception, as mentioned, cannot be limited by the rightholders.

---

[203] For equality as a public interest in this sense, see Weichert, in: Kühling/Buchner (eds.), DS-GVO BDSG, 2nd ed. 2018, Art. 9 DS-GVO para. 90.
[204] Mészáros and Ho, op. cit. *supra* note 193, at 415 et seq.
[205] Buchner/Tinnefeld, in: Kühling/Buchner (eds.), DS-GVO BDSG, 2nd ed. 2018, § 27 BDSG para. 8.
[206] See for a discussion in English Mészáros and Ho, op. cit. *supra* note 193, at 412-414.
[207] See Mészáros and Ho, op. cit. *supra* note 193, at 405; BT-Drucks. 18/11325, 99; see also Buchner/Tinnefeld, in: Kühling/Buchner (eds.), DS-GVO BDSG, 2nd ed. 2018, Art. 89 DS-GVO para. 12 et seq.
[208] On the classical hold-up problem following from sunk costs, see Klein, Crawford, and Alchian, 'Vertical integration, appropriable rents, and the competitive contracting process' 21 *The Journal of Law and Economics* (1978), 297 (301 et seqq.).

This question ought to be answered in the affirmative: commercial research should benefit from the Article 3 exception, too.[209] Such a result could be achieved by an interpretation of the CDSM Directive in conformity with primary EU law (Art. 20 of the Charter). A convincing, legitimate reason for withholding the benefits of the Article 3 exception to research conducted within commercial companies does not seem to exist. To do so not only hurts large corporations, like Google or Facebook, who can more easily transfer their research units to jurisdictions with friendlier research exceptions. It arguably hits most harshly other entities, like journalists or small and medium enterprises (SMEs) in the EU, who do not have this geographical flexibility.[210] However, as the Commission White Paper rightly points out, innovative research at the level of SMEs is one of the backbones of the EU economy; opportunities for AI research within SMEs should generally be fostered, not restricted.[211] This holds particularly in the case of the Article 3 exception which only applies to data to which developers have already gained access legally – and for which, therefore, the rightholders have already had the opportunity to collect remuneration.

Hence, an interpretation of Article 3 CDSM Directive, in the light of Article 20 of the Charter, must disregard the non-binding recital 12(7) CDSM Directive which excludes research units under the decisive influence of commercial companies. Rather, these units must be able to avail themselves of the Article 3 exception if they fulfill the criteria laid down in Article 2(1) CDSM Directive for research organizations in general. Under such an understanding, for example, an AI research unit controlled by a company would qualify if it reinvests all profits into research, even if the research output is also used for product development. This seems to strike a reasonable balance, in conformity with primary EU law, between the commercial interests of the rightholders and the innovation opportunities mentioned in recitals 5 and 8 CDSM Directive, including the research and development interests of companies – which is where, at the moment, considerable advances in AI research take place. Those are advances which many consumers take for granted on a daily basis. Being restricted to a copyright exception, such an understanding does not offer *carte blanche* to AI companies to do whatever they want. Rather, this reading would ensure consistency with data protection law, where Article 89 GDPR (and § 27 BDSG) cover both commercial and non-commercial research, too.

VI. Conclusion: Risk-based technology design through law

The analysis has shown that three risks are crucial for a legal framework for AI training data: data quality, discrimination and innovation risks. The risk of blockage by intellectual property rights has been addressed by Art. 3 et seq. CDSM Directive and, essentially, been solved appropriately. Only the treatment of commercial research must be brought into conformity with the principle of equal treatment under the Charter. However, questions of the quality and non-discriminatory features of training data as well as blockage by existing data protection rights have not yet been sufficiently covered by existing EU law. Overall, the regulatory framework must emancipate itself from the perennially controversial issue of personal identifiability of training data, implied by data protection law, and develop overarching standards.

The regulatory process commenced with the Commission White Paper offers a unique window of opportunity in this respect. A novel EU regulation would be desirable which, irrespective of

[209] Similarly, from a policy perspective, Ducato and Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"', 50 *IIC* (2019), 649 (666); Margoni and Kretschmer, 'The Text and Data Mining exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs', Working Paper, 2018, 4 et seq.; Geiger, Frosio and Bulayenko, op. cit. *supra* note 101, p. 20 et seq.; Obergfell, op. cit. *supra* note 101, p. 223 (230 et seq.).
[210] Cf. Margoni and Kretschmer, op. cit. *supra* note 209.
[211] European Commission, op. cit. *supra* note 1, p. 3, 7.

the applicability of the GDPR, defines standards for the quality and non-discrimination features of training data and training environments according to a risk-based approach. Suggestions for a concrete specification of these criteria were made throughout the article. Simultaneously, in the event that the personal identifiability criterion is met in an individual case, the regulation should contain concrete guidelines for the admissibility of re-using such data as AI training data under data protection law. The instrument would, in this sense, constitute a problem-specific *lex specialis* implementation of the GDPR. In this way, the hitherto quite vague trade-offs of Article 6(1)(f) GDPR, Article 6(4) GDPR and national transpositions of Articles 9(2)(j), 89 GDPR (e.g., § 27(1) BDSG) could be operationalized in a context-specific manner.

Overall, a legal framework for training data affords the advantage of actively shaping AI applications ex ante, at the stage of their technical design, in such a way that elementary legal norms and social values are respected. In contrast to human decisions that can hardly be explicitly controlled, this possibility, of consciously determining the relevant parameters, also demonstrates the considerable promise of responsible AI for socially desirable decisions. Thus, in the field of AI training data, scholars currently have the rare opportunity to carry out interdisciplinary, basic legal research which, simultaneously, critically accompanies ‘in real time’ one of the central regulatory projects of our time: the EU Commission's plans for regulating AI. The present contribution offers first steps in the direction of such an endeavor.