Regulating AI with Purpose Limitation for Models

*Rainer Mühlhoff and Hannah Ruschemeier**

This article proposes the concept of purpose limitation for AI models as an approach to effectively regulate AI. Unregulated (secondary) use of specific models creates immense individual and societal risks, including discrimination against individuals or groups, infringement of fundamental rights, or distortion of democracy through misinformation. We argue that possession of trained models, which in many cases consist of anonymous data (even if the training data contains personal data), is at the core of an increasing asymmetry of informational power between data companies and society. Combining ethical and legal aspects in our interdisciplinary approach, we identify the trained model, rather than the training data, as the object of regulatory intervention. This altered focus adds to existing data protection laws and the proposed Artificial Intelligence Act. These are inefficient in preventing the misuse of trained models due to their focus on the procedural aspects of personal data or training data. Drawing on the concept of risk prevention law and the principle of proportionality, we argue that the potential use of trained models by powerful actors in ways that are damaging to society warrants preventive regulatory interventions. Thus, we seek to balance the asymmetry of power by enabling democratic control over where and how predictive and generative AI capabilities may be used and reused.

Keywords: EU AI Act; GDPR; purpose limitation; regulating models; data power

I. Introduction

Artificial intelligence (AI) technology plays an important role in numerous application areas today. Most socially relevant use cases of AI rely on machine learning techniques. These are algorithms that are configured ('trained') based on vast amounts of data to identify 'patterns'. Subsequently, they can be used to recognise patterns in input data, which forms the basis for their output. For example, when machine learning models are used in predictive analytics, the output data contains risk scores or predictions of unknown characteristics such as health dispositions, sexual orientation, religious and ethnic belonging, political views, educational background, and financial status.[1] When machine learning models are used in natural language processing or image recognition, the pattern detection is linked to text production, such as when a transcription is produced for an identified word in audio data or a label is produced for an identified object in an image. In generative AI, pattern detection is combined with the extrapolation of these patterns, eg, ChatGPT extending the input prompt with the most likely answer as output.

All of these diverse machine learning applications have two structural aspects in common: first, they

DOI: 10.21552/aire/2024/1/5

* Rainer Mühlhoff, Full Professor of Ethics of Artificial Intelligence at the University of Osnabrück, Germany. Hannah Ruschemeier, Junior Professor (tenure W3) for Public Law, Data Protection Law and Law of Digitalisation at the University of Hagen, Germany. For correspondence: <rainer.muehlhoff@uni-osnabrueck.de> and <hannah.ruschemeier@fernuni-hagen.de>. All internet links were last accessed 19 February 2024.

1 See, eg, Hans Lammerant and Paul de Hert, 'Predictive Profiling and Its Legal Limits: Effectiveness Gone Forever' in B van der Sloot, D Broeders and E Schrijvers (eds), *Exploring the boundaries of big data*, vol 32 (Amsterdam University Press/WRR 2016); Rainer Mühlhoff, 'Predictive Privacy: Collective Data Protection in the Context of AI and Big Data' (2023) 10 Big Data & Society 205395172311668; Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008); Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019 Columbia Business Law Review 494.

rely on the training data obtained from often thousands to millions of individuals and across different sources such as the users of digital services. Second, providers can often reuse the trained models without substantial legal barriers for numerous secondary purposes, including risky and malicious ones, ranging from differential pricing to fake news. In many cases, a trained model constitutes a set of highly aggregated, anonymous, and, therefore, non-personal data that can be freely circulated, sold, and published without data protection hurdles.

Our article contends that the lack of regulation of trained models presents a severe threat to individuals and society that urgently needs regulatory attention. Trained models are powerful tools as they can be used or reused for automated decisions, behavioural scoring, or discriminatory business practices. These include using models that can predict psychological character traits in targeted political advertisements, models able to predict the prevalence of medical conditions based on social media data being reused in the insurance industry as well as medical models that can predict substance abuse or psychological dispositions such as depression being reused in AI-assisted hiring procedures.[2] Likewise, the risk of abusing generative AI models is abundant. It ranges from the production of fake personal statements that interfere with personal rights, photos or videos which violate copyrights, to the production of fake evidence in news images, or text that creates hate speech and fake news.[3]

In this article, we introduce *purpose limitation for models* as the conceptual idea of a regulatory approach to this unresolved problem. Purpose limitation for models demands that the production and use of AI models must be limited to specific purposes. These must be stated *ex ante* and enforced throughout the life cycle of an AI model. Analogous to purpose limitation in the processing of personal data familiar from data protection, purpose limitation for models strives for a state in which trained models are not allowed to be used for any other purpose than their primary ones and must be deleted when they are not needed for that purpose anymore. The detailed elaboration of a positive list of admissible purposes and the underlying ethical principles to validate purposes of AI models is the subject of a separate piece of work to be conducted with participatory methodology and stakeholder involvement. In this article, our aim is to introduce the theoretical framework of a regulatory approach that rearticulates the concept of purpose limitation not for training data but for trained models. Given that purpose limitation is a fundamental principle of European Data Protection Law,[4] it has been comparatively overlooked in the governance of AI models. Attention has primarily been paid to transparency and fairness. However, in contrast to purpose limitation in the processing of personal data, the entity empowered by purpose limitation for models cannot be the individuals whose data appear in the training data. Rather, the entity that is empowered to make decisions about valid purposes must be an agent that can consider collective impacts and collective interests such as an oversight body under democratic control.

On the one hand, there is currently no AI governance of the purposes for which AI models may be built and used. The decisions about whether, how, and by whom these models are developed and used are entirely in the hands of a few globally operating economic players – the Big Tech companies. On the other hand, the implications of how trained models are used and reused typically affect large numbers of people, collective decision-making processes such as elections, and entire societies. This establishes a power shift at the benefit of private AI companies. We refer to the unilateral ability of large data and AI companies to train and use predictive and generative AI models as a new manifestation of informational power asymmetry between data processing entities and societies. Purpose limitation for models is a regulatory proposal that aims directly at the transformative effects of AI on power relations.[5] To balance these new informational power asymmetries, the decision-

2 See for examples of wrongful secondary use: Rainer Mühlhoff, 'Das Risiko Der Sekundärnutzung Trainierter Modelle Als Zentrales Problem von Datenschutz Und KI-Regulierung Im Medizinbereich' in Hannah Ruschemeier and Björn Steinrötter (eds), *Der Einsatz von KI & Robotik in der Medizin* (Nomos 2024) <https://www.nomos-elibrary.de/10.5771/9783748939726-27/das-risiko-der-sekundaernutzung-trainierter-modelle-als-zentrales-problem-von-datenschutz-und-ki-regulierung-im-medizinbereich?page=1>.

3 Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and Other Large Generative AI Models', *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023) <https://dl.acm.org/doi/10.1145/3593013.3594067>.

4 On this, eg, Merel Elize Koning, 'The Purpose and Limitations of Purpose Limitation' (Radboud University Nijmegen, 2020) <https://repository.ubn.ru.nl/bitstream/handle/2066/221665/221665.pdf?sequence=1>.

5 Pratyusha Kalluri, 'Don't Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power' (2020) 583 Nature 169.

making processes underlying the use of trained models need to change away from the actors' economic interests.

In this article, we take an approach that tackles power asymmetries between private actors (industry)[6] and society via effective regulation. Discussions around AI often emphasise the importance of making AI technology fair, transparent, widely available, affordable, and empowering for end users. Here, we focus on the contexts and purposes for which AI is deployed. In these contexts, AI acts as a lever, exacerbating existing inequalities and intensifying coercive forces. One of the factors that contributes to the unaccounted multiplication of informational power asymmetry is the ability to legally reuse trained AI models for secondary purposes that might not be publicly visible. We therefore need an approach that does not only focus on the *intended* use of a system but on the potential for later reuse. This requires consideration of the legal, economic, societal position of the actors, the potential dissemination of trained models, the potentially affected legal interests and the larger collective effects of these models beyond their original context of application. Purpose limitation for models thus aims to implement a loop of democratic oversight and control at the level of building, using and reusing trained models.

Producing models is not a neutral step from an ethical, political, and legal perspective. Rather, possession of a trained model implies a strong form of informational power. We analytically distinguish the production of a model ('training') from its subsequent application on a specific individual or case (see Section II.1). Existing data protection regulation only protects against abuse when personal data is processed. We argue that a trained model itself – which in general consists of non-personal data – needs regulatory attention. This is because the mere existence of a trained model that can be freely circulated and repurposed poses severe societal risks. Under current legislation, neither the data subjects whose data is being used as training data nor society as a whole can effectively control whether AI models are being built from their data and how these models are (re)used. As we will argue, equipping the individual data subject with means of control does not resolve the problem as only extensive collections of data, not individual data points, enable the training of machine learning models. In this inherently collective structure that enables machine learning technology, collective control and regulation structures are needed.

In Section II, we argue why and in which ways the uncontrolled existence of trained models poses a risk to society, collective interests, and individual fundamental rights. We will refer to the concept of risk prevention law to argue our case for better regulation. In Section III, we introduce purpose limitation for models as a concept. After arguing for this strategy on ethical grounds, we then discuss why current data protection regimes are insufficient to control trained models. In Section IV, we outline the first steps of a regulatory proposal of purpose limitation for models that addresses the collective structure of machine learning models.

II. What Is the Problem?

1. Data Processing Chain

Our proposal for a purpose limitation for models is closely related to the data processing life cycle of machine learning systems. In order to precisely identify where and how purpose limitation shall apply, we distinguish three steps in the typical data processing chain of such systems: (1) training data collection, (2) model training, and (3) model application.

(1) Training Data Collection: In the first step of creating a machine learning system, many data points are collected as training data.[7] These data points can include personal or non-personal data. In some cases, extremely large data sets are used for this step. This makes it difficult, if not impossible, to distinguish between different categories of data. ChatGPT, for example, was trained from vast amounts of data that are freely accessible on the web.[8]

(2) Model Training: As a second step, a machine learning model is trained on the collected training

6 The situation does not improve when state actors use predictive models as they often cooperate with private actors and the state-citizen relationship increases the power asymmetry to the same extent.

7 Cf Kevin P Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press 2012) 1; Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st edn, Crown 2016).

8 Tom Brown et al, 'Language Models Are Few-Shot Learners' in H Larochelle et al (eds), *Advances in neural information processing systems* (Curran Associates, Inc 2020) <https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>.

data. This training procedure is an algorithm that seeks to 'learn' correlations, patterns or any other information from the training data trove, resulting in a configured ('trained') model.[9] Such a model could be a trained Artificial Neural Network (ANN) or other implementation of machine learning algorithms.[10]

(3) Model Application: In the third step in the typical data processing chain, the trained machine learning model is applied to specific cases or individuals. This means that the model is used as a tool that computes a specific output in response to input data. In this step, the model output constitutes information about new cases or third parties. These can potentially apply to individuals or cases that were not part of the training data.[11] In generative AI systems, some prompt is fed to the model as input, which generates, for instance, text or an image as output. In the case of classifiers or predictive models, data that is available about the case at hand is input for the model. This could be the CV of a job applicant that is inputted to a model that assists in the shortlisting of job applications or the social media data of an Instagram user that is inputted to a model that can predict the user's current emotional state. The application of the model does not need to follow the model training immediately. Rather, the model can be used much later or by different entities that gain access to the trained model. To apply the model, access to the training data is not needed.

2. The Risk of Secondary Use

In this paper, we address the risk of trained models originating from step 2 to be reused or repurposed for applications that are harmful to individuals or society.[12] Imagine a social media platform builds a model that can predict alcohol consumption from users' behavioural data (eg, 'liked' items and visited websites). Its initial purpose is to serve ads and relevant content to users in their news feed.[13] The kind of unaccounted secondary use we address in this paper refers to instances when such a model gets repurposed, eg, in algorithmic hiring systems, where a copy of the model is used for screening job applicants.

The risk of unaccounted, and often harmful, secondary use of a trained model can easily evade public attention. This is especially the case when the model was originally created for a presumably beneficiary purpose. In the typical scenario we have in mind, the original and publicly communicated purpose for adopting machine learning technology is at a minimum uncontroversial. The public, economic, and political attention often focuses on the opportunities for innovation it provides. In such situations, the dangerous reuse of trained models is often not acknowledged in public debates surrounding the respective technological innovation. This repurposing can also occur years later and involve different actors (eg, different companies after a merger or acquisition of the original company). While the next subsection (II.3) details what we mean by the harmful use of AI models, we outline why there is a real risk that trained models get transferred to potentially unforeseen secondary use cases in the following.

It is central to our argument that the trained model which originates from processing step 2 constitutes data in its own right, distinct from the training data.[14] We refer to this data as *model data*. The model data represents the trained state of the model. A trained ANN, for instance, is represented by a large matrix of numbers determined by the weights and other parameters (eg, activation thresholds) of the 'neurons' and

9 The use of terms such as 'training' and 'learning' has been criticised as anthropomorphising AI systems. Stating that a model is 'configured' instead of 'trained' avoids this pitfall but comes with the disadvantage that this terminology is less popular. See Rainer Rehak, 'The Language Labyrinth: Constructive Critique on the Terminology Used in the AI Discourse' in Pieter Verdegem (ed), *AI for Everyone? Critical Perspectives* (University of Westminster Press 2021).

10 See on the different tasks and different types of algorithms Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* (The MIT Press 2016) 99 et seq.

11 Rainer Mühlhoff, 'Predictive Privacy: Towards an Applied Ethics of Data Analytics' (2021) 23 Ethics and Information Technology 675; Rainer Mühlhoff and Hannah Ruschemeier, 'Predictive Analytics and the Collective Dimensions of Data Protection' (2024) 16(1) Law, Innovation and Technology <https://www.tandfonline.com/doi/full/10.1080/17579961.2024.2313794>.

12 Mühlhoff (n 2).

13 That the prediction of substance abuse and many other psychosocial, health-related issues is possible based on social media data is well established: Michal Kosinski, David Stillwell and Thore Graepel, 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior' (2013) 110 Proceedings of the National Academy of Sciences 5802; Raina M Merchant et al, 'Evaluating the Predictability of Medical Conditions from Social Media Posts' (2019) 14 PLOS ONE e0215476. The risks of the reuse of models that were originally trained for targeted advertising purposes is specifically debated in Rainer Mühlhoff and Theresa Willem, 'Social Media Advertising for Clinical Studies: Ethical and Data Protection Implications of Online Targeting' [2023] Big Data & Society <https://journals.sagepub.com/doi/epdf/10.1177/20539517231156127>.

14 Mehtab Khan and Alex Hanna, 'The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability' (13 September 2022) <https://papers.ssrn.com/abstract=4217148>.

their connections.[15] Other machine learning models have other ways of representing their internal parameters as data. To evaluate which restrictions concerning the legal processing of model data are in place, it is of interest whether the trained model (model data) can be classified as personal data. There is no universal answer to this question. If the training data is personal data, the model data can be either personal or anonymous data, depending on the training procedure. For instance, if state-of-the-art anonymisation techniques such as differential privacy and federated machine learning are used, it is in theory possible that a model is produced in step 2 that does not contain any back references to the training data.[16] In this case, legal restrictions on the processing of personal data would not apply to the model data. Importantly, the potential for harmful applications of a trained model (see II.3) does not diminish once model data is anonymous. Hence, in order to envision the most severe regulatory gap, it is reasonable to assume that the model data is anonymous, even if the training data contains personal data.

Hence, while we assume that the model data itself is typically anonymous data and, therefore, does not fall within the scope of the General Data Protection Regulation (GDPR),[17] we are potentially moving back into the realm of personal data in processing step 3. This is because in the typical situation we have in mind, the input data in the application stage is data linked to a specific person or case and the output constitutes data about this person or case (eg, a prediction, a classification or a generated text or image that relates to the person or case). This person or case to which the model is applied does *not* need to be part of the training data that was used in step 2 to produce the model.[18] To illustrate, a model to predict alcohol abuse from social media data trained on the (anonymised) data of individuals 1–1000 can be applied to the behavioural data of user number 1001 to predict their likelihood of substance abuse. Note that the data processing chain involves two different types of data subjects: the data subjects of the training data set and the data subjects of the application step.[19] In step 2 referring to the storing, using, and potential repurposing of a trained model, we generally have no data subjects as the model data is highly aggregated and, in many cases, even anonymous.

We argue that existing legislation related to individual data subjects in processing steps 1 and 3 is insufficient to prevent harmful secondary use of model data originating from step 2. Model data that represents trained models is currently not governed by specific regulation and evades the focus of existing regulation that hinges on the categorisation as personal data. Still, this data poses a considerable risk to society if it can be freely processed, including its sale and circulation. This is because the trained model has the potential to be applied for *any* purpose and to *any* individual or case – present, past, and future – singly or in parallel (through mass processing) in ways that are beyond the reasonable control of democratic policy. Therefore, we propose introducing a purpose limitation for models between steps 2 and 3 in section III to prevent *ex ante* risky applications that have an adverse impact on individuals and society.

3. Societal Risks and Dangers Connected to AI Models

The spectrum of individual and societal risks connected to the application of big data and machine learning is broad. It includes concerns about increasing social injustice,[20] opaque bias and discrimination in financial, hiring or welfare decisions,[21] new forms of privacy violation,[22] capitalist and colonialist exploita-

15 Goodfellow, Bengio and Courville (n 10).

16 Martín Abadi et al, 'Deep Learning with Differential Privacy' [2016] Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16 308; Cynthia Dwork, 'Differential Privacy' in Michele Bugliesi et al (eds), *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*, vol 2 (Springer 2006).

17 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

18 Mühlhoff (n 11).

19 Khan and Hanna (n 14).

20 O'Neil (n 7); Pieter Verdegem (ed), *AI for Everyone? Critical Perspectives* (University of Westminster Press 2021) <https://www.uwestminsterpress.co.uk/site/books/e/10.16997/book55/>; Rainer Mühlhoff, 'Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus' (2020) 68 Deutsche Zeitschrift für Philosophie 867.

21 Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671; Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (1st edn, St Martin's Press 2017); Sandra Wachter, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law' (2023) 97 Tulane Law Review 149.

22 Mühlhoff (n 11); Mühlhoff (n 1); Mühlhoff and Ruschemeier (n 11); Wachter and Mittelstadt (n 1); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (Profile Books 2019).

tion of human and planetary resources,[23] threats to democracy from disinformation,[24] and infringement of copyright and personal rights[25] amongst others.

For the purposes of this paper, we cannot extensively review these diverse debates. Instead, we highlight two (out of several) categories of potential abuse related to trained machine learning models that will serve as illustrating examples throughout this paper. The first one is predictive analytics and concerns the use of machine learning models to predict unknown information about individuals or cases. The example of predicting alcohol consumption from social media data that has been mentioned in II.2 falls into this category. More generally, it has been shown that various addictions and diseases, including substance abuse, depression, psychosis, diabetes, and high blood pressure, can be predicted from social media data.[26] These methods are controversial, however. Insurance and finance companies are interested in these predictive models as they enable individual risk assessment beyond traditional credit scores.[27] In these domains, as well as in areas such as human resource management, predictive models could lead to implicit discrimination based on sensitive attributes like race or pregnancy.[28] Predictive analytics is further widely used in targeted advertising. Here, exploiting real-time information about users' vulnerabilities and emotions potentially leads to manipulative practices. These practices – sometimes debated as 'hypernudges'[29] or 'dark patterns' – combine prediction and manipulation. This in turn poses risks to user autonomy as exemplified by Facebook's targeting of emotionally vulnerable teenagers with specific advertisements.[30]

Treating individuals differently based on predictively modelled traits and behaviour undermines democratic principles as it leads to stereotypes, mistreatment of outliers, and epistemic injustice.[31] Preventive protection of individual rights, collective interests, and supra-individual government processes is therefore necessary in order for society to be able to control the risks emerging from predictive models. This is in particular the case when these models are reused beyond their original purposes. In democratic political systems, collective decision-making processes presuppose the autonomy of the individual.[32] This autonomy is diminished by algorithmically generated attributions over which individuals have no control. In addition, applying predictive models in different contexts poses risks to the rights to privacy and non-discrimination. Epistemically, the transition to a prediction-based knowledge order based on correlations rather than causalities is problematic due to the lack of quality assurance mechanisms.[33]

A second category of potential abuse we would like to highlight concerns generative AI, specifically, the risk of generative models being used for the production of false evidence and news reports[34] or deepfake imagery.[35] It has been shown that individuals are 'largely incapable of distinguishing between AI- and human-generated text'.[36] The dangers associat-

23 Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press 2021); Danielle Coleman, 'Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws' (2018) 24 Michigan Journal of Race & Law 417; Jathan Sadowski, *Too Smart: How Digital Capitalism Is Extracting Data, Controlling Our Lives, and Taking over the World* (MIT Press 2020).

24 Zeynep Tufekci, 'Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency' (2015) 13 Colorado Technology Law Journal 203; Hacker, Engel and Mauer (n 3).

25 Matthew Sag, 'Copyright Safety for Generative AI' (4 May 2023) <https://papers.ssrn.com/abstract=4438593>.

26 Merchant et al (n 13); Mühlhoff and Willem (n 13).

27 O'Neil (n 7) ch 8.

28 Ibid, 108, 148.

29 Karen Yeung, ''Hypernudge': Big Data as a Mode of Regulation by Design' (2017) 20 Information, Communication & Society 118.

30 Daniel Susser, Beate Roessler and Helen Nissenbaum, 'Online Manipulation: Hidden Influences in a Digital World' (2019) 4 Georgetown Law Technology Review 1; Tal Z Zarsky, 'Privacy and Manipulation in the Digital Age' (2019) 20 Theoretical Inquiries in Law 157.

31 Dan McQuillan, 'Predicted Benefits, Proven Harms: How AI's Algorithmic Violence Emerged from Our Own Social Matrix' [2023] The Sociological Review Magazine <https://thesociologicalreview.org/magazine/june-2023/artificial-intelligence/predicted-benefits-proven-harms/>; Justin Joque, *Revolutionary Mathematics: Artificial Intelligence, Statistics and the Logic of Capitalism* (Verso 2022); Mühlhoff (n 11).

32 BVerfG, Order of the First Senate of 15 December 1983 – 1 BvR 209/83. English version: <https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/1983/12/rs19831215_1bvr020983en.html>.

33 Joque (n 31); Rainer Mühlhoff, *Die Macht der Daten: Warum künstliche Intelligenz eine Frage der Ethik ist* (1st edn, V&R unipress 2023) <https://www.vr-elibrary.de/doi/book/10.14220/9783737015523> accessed 11 May 2023.

34 Ben Buchanan et al, 'Truth, Lies, and Automation: How Language Models Could Change Disinformation' (Center for Security and Emerging Technology 2021) <https://cset.georgetown.edu/publication/truth-lies-and-automation/>.

35 Don Fallis, 'The Epistemic Threat of Deepfakes' (2021) 34 Philosophy & Technology 623.

36 Sarah Kreps, R Miles McCain and Miles Brundage, 'All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation' (2022) 9 Journal of Experimental Political Science 104.

ed with this technology multiply considerably due to the easy scalability of AI systems that could result in the amplification of minority perspectives to produce a distorted version of the majority discourse.[37] This can ultimately lead to voter manipulation and the spread of misinformation that influences real-life processes and deteriorates democracy.[38]

4. Insufficient Regulation of Trained Models *De Lege Lata:* GDPR and Anti-Discrimination Law

The principle of purpose limitation in data processing is a cornerstone of the GDPR laid down in the provision about principles of Art 5. It mandates that data controllers must define the purpose of data collection no later than at the point of collection and restricts them from processing the data in any manner that diverges from the initially stated purpose as stipulated in Article 5 (1) (b). These purposes need to be specific, explicit, and legitimate in order to define the aim and goal of the data processing. Therefore, the purpose limitation principle is closely related to the principles of storage limitation and data minimisation.[39] The principle does not strictly bind data processing to the original purpose. Rather, the secondary data use has to be *compatible* with the original purpose (cf Article 5 (b)). Compatibility is concretised by two specifications: First, according to Article 5 (b), the privileged processing purposes (for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes) mentioned there are considered compatible with the initial purpose in the sense of a legal fiction in accordance with Article 89 (1). Second, Article 6 (4) GDPR formulates a compatibility test and offers a series of criteria to determine whether the processing for a purpose other than the one for which the personal data has been collected is to be considered compatible with its initial purpose.

Purpose limitation must be explicitly defined at the start of data processing. As a consequence, the data processor is required to specify the purposes of the data processing as a first step. This is also a prerequisite for other GDPR requirements such as data minimisation. The second step requires an examination of whether the further processing is a privileged purpose under Article 5 (b), 89 (1) GDPR. If this is not the case, a subsequent third step necessitates assessing the requirements of Article 6 (4) GPDR. These postulate that the requirement of the compatibility test does not apply if the data processing is a) based on consent or b) on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23 (1). In all other cases, the data processor must check the compatibility of the purposes according to the criteria outlined in Article 6 (4).

The aim of purpose limitation is to enable data subjects to make informed choices about which actors process their data and for which purposes.[40] The purpose limitation principle considers that once data is collected and stored, it could be used for any purpose. This could potentially infringe the data subject's right to the protection of personal data. Additionally, the purposes pursued must be legitimate, meaning that they must follow not only data protection law but the more comprehensive legal order. The intent here is not to burden data subjects with the responsibility of verifying the legitimacy of these purposes. Instead, this responsibility squarely lies with the data processors. In fact, the main goal of purpose limitation is to protect the data subject and to enable the controllability of further data processing and its compliance with data protection law.

Purpose limitation is 'old data protection law'.[41] With its roots in Article 8 of the European Charter of Fundamental Rights, it has been a core principle of data protection law even since before the GDPR came into force. It has been claimed that, when it comes to AI and Big Data Technologies, already the purpose specification, let alone its limitation, seems difficult to execute.[42] As with all data protection principles, there are significant enforcement deficits regarding purpose limitation ('*enormous disconnect between*

37 Buchanan et al (n 34).

38 Jiawei Zhou et al, 'Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (ACM 2023) <https://dl.acm.org/doi/10.1145/3544548.3581318>; Brahim Zarouali et al, 'Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media' (2022) 49 Communication Research 1066.

39 Michele Finck and Asia J Biega, 'Reviving Purpose Limitation and Data Minimisation in Data-Driven Systems' (2021) 2021 Technology and Regulation 44.

40 Article 29 Data Protection Working Party, WP203/569/13 (2013).

41 Finck and Biega (n 39).

42 Mireille Hildebrandt, 'Slaves to Big Data. Or Are We?' (2013) 17 Ipd. revista de internet, derecho y política 7, 35. 7.

law and reality').[43] Moreover, this enforcement deficit is not substantially mitigated by the numerous fine proceedings and recent case law from the ECJ. These decisions have not led to a change in data-invasive business models, which in our view are not compatible with the GDPR.[44] Powerful actors process data for hundreds of vague and unspecified purposes.[45] A standard practice among these entities is to collect data and subsequently define its uses.

We argue that the challenge described cannot even be resolved by more effective enforcement of existing data protection law alone, although this is necessary.[46] Current legislation does not address purpose limitation for models sufficiently for three reasons: First, models consisting of anonymised data do not fall within the scope of the GDPR which only addresses the processing of personal data. The assumption that anonymisation itself can be the data processing subject to authorise does not resolve this issue as any purpose limitation or restriction is lost after anonymisation.

Second, the GDPR's assumptions about data processing operations often no longer align with reality. In its recent ruling in the Meta case,[47] the ECJ acknowledged that the distinction between personal and non-personal data becomes *de facto* obsolete.[48] Respectively, the example of ChatGPT illustrates that when large troves of data are scraped from the Internet, it is no longer possible to differentiate the database *ex post* with respect to normative categories.[49] If the object of regulation of the processing of personal data becomes increasingly challenging to identify, mechanisms such as purpose limitation, understood as individual protection, can no longer be effective. Notably, these problems are intensified by the most widespread legal basis for the processing of personal data in practice: consent. Many authors, including ourselves, have argued that consent is an unsuitable legal instrument for the legitimation of data processing in digital environments.[50] The ECJ did not assume that voluntary consent was excluded merely due to the dominant position of a social media platform like Meta.[51] In this specific situation, this is understandable in the context of competition law. The problem of consent in the digital environment is not only due to the market position of an actor but also due to the sheer flood of information.[52] Yet, when it comes to the informational power asymmetries that result from machine learning models, the same actors have created business models that are impossible for individuals to oversee and understand. These include constellations in which consent is given to 300 different data processors simultaneously. Additionally, predictions are new personal data that the data subject cannot foresee while consenting at the time of data processing.

Third, even for models that process personal data, cases of secondary data use through resale are not effectively regulated.[53] In many cases of secondary data use, the purpose limitation principle is no longer traceable. Even if data protection policies are publicly available and define purposes like 'personalis[ing] content' or 'improv[ing] services,' there is, in fact, no control over whether this corresponds to the actual data processing practice.[54] Among other things, this is because a distinction must be made between cases where the purpose is changed by the same data processor and further use by third parties. In the first

43 Bert-Jaap Koops, 'The Trouble with European Data Protection Law' (2014) 4 International Data Privacy Law 250, 256.

44 Mühlhoff and Ruschemeier (n 11).

45 Isabel Hahn, 'Purpose Limitation in the Time of Data Power: Is There a Way Forward?' (2021) 7 European Data Protection Law Review 31, 41.

46 Ibid.

47 C-252/21 *Meta v Bundeskartellamt* [2023] OJ 62021CJ0252.

48 Cf. for the criticism of the category itself Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40.

49 Hannah Ruschemeier, 'Squaring the Circle' (*Verfassungsblog*, 7 April 2023) <https://verfassungsblog.de/squaring-the-circle/>.

50 Omri Ben-Shahar and Carl E Schneider, 'The Failure of Mandated Disclosure' (2011) 159 University of Pennsylvania Law Review 647; Frederik J Zuiderveen Borgesius et al, 'Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the ePrivacy Regulation' (2017) 3 European Data Protection Law Review 353; Sourya Joyee De and Abdessamad Imine, 'Consent for Targeted Advertising: The Case of Facebook' (2020) 35 AI & SOCIETY 1055; Trung Tin Nguyen, Michael Backes and Ben Stock, 'Freely Given Consent? Studying Consent Notice of Third-Party Tracking and Its Violations of GDPR in Android Apps', *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery 2022) <https://dl.acm.org/doi/10.1145/3548606.3560564>; H Brian Holland, 'Privacy Paradox 2.0' [2010] Widener Law Journal 883; Hannah Ruschemeier, 'Privacy Als Paradox?' in Michael Friedewald et al (eds), *Künstliche Intelligenz, Demokratie und Privatheit* (Nomos 2022).

51 C-252/21, paras 140-141.

52 Ruschemeier (n 50).

53 Hannah Ruschemeier, 'Data Brokers and European Digital Legislation' (2023) 9 European Data Protection Law Review 27.

54 Finck and Biega (n 39), 50.

case, the requirements for a purpose change have to be determined according to Articles 5 (1) (b), 6 (4) GDPR. In contrast, the second case including, eg, the sale of data sets, is considered new data processing by a third-party controller. This new processing can be based on a new legal basis in line with Article 6 without being subject to the restrictions of the original purpose limitation. Further processing then only has to be compatible with the new purpose(s).[55] According to the current understanding, anonymisation or pseudonymisation should always be compatible as the risks for data subjects are considered low here. In any case, incompatible purposes can be overcome by the data subject's consent, which is not an adequate safeguard due to the inappropriateness of consent. At first sight, it is unclear whether the function of Article 6 (4) is limited to a compatibility test or whether Article 6 (4) is also to be classified as an authorisation for further processing of personal data for another purpose. But according to the wording and the systematology, Article 6 (4) can only refer to the interpretation of the requirement of compatibility under Article 5 (1) (b) since it is a question of compatibility and cannot provide for an exception to the general rule of Article 6 (1). Indeed, Article 6 (1) refers only to letters a to f and not to paragraph 4 of the provision. Therefore, the requirements from Article 6 (4) specify Article 5 (1) (b) GDPR.

Even if national or EU anti-discrimination law[56] were to apply to steps 2 and 3 of the data processing chain, it would suffer from the same enforcement deficits as data protection law.[57] Due to the collective dimensions of machine learning, these cases are a form of 'victimless discrimination:' The law assumes that individual data subjects will assert their rights, but they are no longer identifiable, and even if they were, the barriers to enforcement are too high due to the power asymmetries described.

5. Legislative Framework for Anonymous Data

The current legal framework for anonymous data does not address the aspect of informal power asymmetries and prediction power. Rather, legislation so far follows a dichotomy between the protection of individuals through data protection law and the protection of non-individual goods, such as the free flow of data to support the single market economy. Especially the regulation on a framework for the free flow of non-personal data in the EU (2018/1807) pursues the goal of developing the data economy and enhancing the competitiveness of the European Union's industry. Hence, it needs to address the issues of the data-powerful actors which we discuss here. The regulation does not refer to any effect on (non-professional) users since it relates to data localisation requirements, the availability of data to competent authorities, and the porting of data for professional users (cf Article 1). Since any regulation of non-personal data follows the dichotomy between personal and non-personal data, it becomes obsolete in massive data sets or large language models. Even though the regulation explicitly mentions that it should only apply to the non-personal data part of a mixed data set and should not prejudice the GDPR (Article 2 (2)), this seems complicated to distinguish in practice. Providers of a product such as ChatGPT will not be able to differentiate due to the immense databases. The same considerations apply to identifying special categories of personal data.

6. Regulating Training Data

Regulation of AI training data beyond the GDPR only addresses part of the problem described here if at all. We argue that regulation should start with the

55 Article 29 Data Protection Working Party, WP203/569/13 (2013).

56 Eg, the German General Act on Equal Treatment (AGG) of 14 August 2006 (Federal Law Gazette I, p. 1897) as last amended by art 4 of the Act of 19 December 2022 (Federal Law Gazette I, p. 2510); Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L 180/22; Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] L 303/16; Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L 204, 23; Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L 373/37. Further on the problems of AI and anti-discrimination law: Janneke Gerards and Frederik Zuiderveen Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence Articles and Essays' (2022) 20 Colorado Technology Law Journal 1; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI' (2021) 41 Computer Law & Security Review 105567.

57 Philipp Hacker, 'A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act' (2021) 13 Law, Innovation and Technology 257.

models themselves and their potential context of use. Regulating only data, not its context of use, has not proven very successful. Therefore, we argue that the risks associated with training data (eg, quality risks, discrimination risks) materialise in the application of the models to unknown and unlimited purposes. The AI Act regulation (AIA)[58] proposes requirements for training, validation, and testing data in Article 10 for 'high-risk' systems only. It requests that this data 'shall be relevant, representative, and to the best extent possible, free of errors and complete' in view of the intended purpose (cf Article 10 (3)).[59] All of this does not, however, address the risk that a trained AI model, even if it was built from relevant, representative, and 'unbiased' training data, could be used or reused for societally risky and damaging purposes. For example, when training a model to detect malicious skin lesions on skin photos, the quality and representativeness of the training data are crucial issues in obtaining a system that works equally well on all skin types.[60] While such a system could, in its primary use, be a beneficial tool in medical care, the risk of reusing this system for discriminatory purposes, for instance, in insurance risk assessment, only pertains to the context of use and not to the quality of the training data.

III. Purpose Limitation for Models

In the preceding section, we discussed that unregulated AI models could come with significant societal risks for which the current legislative acts do not provide sufficient handling. This section introduces the principle of purpose limitation for models as part of a solution.

1. The General Idea

From an ethical and conceptual perspective, we seek a preventive regulation that puts reasonable limits to the open-ended possibilities of using and reusing trained machine learning models in ways that could be harmful to society. We identify the *model data* – the trained model represented by a data set that originates from step 2 of the typical data processing chain outlined in II.1 – as a regulatory intervention point. From there, we propose implementing a principle of purpose limitation that applies to the processing (eg, storage, circulation, and utilisation) of this model data. As previously established, the model data is distinct from the training data of a model. Therefore, proposing purpose limitation for the model data is different from proposing purpose limitation for the training data as the training data is used only once (in step 1) to produce the trained model and can then be discarded.

In the typical situation we aim to address (see II.2), a model has been built from training data for a purpose that is, in the best case, agreed upon to be beneficial. Yet, once the model is in place, we must assume that the model data consists of anonymous data that does not fall within the scope of existing data protection regulation. It can therefore be repurposed in uncontrolled, including harmful, ways. As this reuse is entirely at the discretion of the model owners (primarily large tech companies), repurposing is currently a unilateral decision. This practice needs to be more exposed to public scrutiny and control as it intensifies an essential aspect of the informational power asymmetry between data and AI companies on the one hand and individuals and society on the other. Regulating this power asymmetry forms the core of our intent.

Crucially, this power asymmetry cannot be mitigated by a regulatory approach that applies only to processing step 1 – the processing of training data – or only to processing step 3 – the application of the model to an individual case (see II.1). To illustrate, take the example of a model that can predict the risk of hepatitis B from social media usage data.[61] Now

58 Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts' COM (2021) 206 final; Council, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach' 15698/22; European Parliament, 'Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))' P9_TA(2023) 0236.

59 Similar in the Commission's proposal and the EP amendments.

60 Lisa N Guo et al, 'Bias in, Bias out: Underreporting and Underrepresentation of Diverse Skin Types in Machine Learning Research for Skin Cancer Detection—A Scoping Review' (2022) 87 Journal of the American Academy of Dermatology 157; Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Conference on Fairness, Accountability and Transparency* (PMLR 2018) <http://proceedings.mlr.press/v81/buolamwini18a.html>.

61 Mühlhoff and Willem (n 13).

imagine this model is covertly repurposed to become part of a system that assists in hiring decisions. The phenomenon that needs controlling is already the existence of the hiring system augmented by the original mode itself. It is at this stage that informational power asymmetry gets inscribed into technology. Any regulation that only controls the application of this system to individual cases misses the actual preventative approach central to controlling power asymmetry. In the case of using predictive models for the selection of job applicants, the necessity of a preventative approach is particularly apparent as there will be no reasonable way for applicants to not consent to this form of data processing.

In addition, it is typical for the situation we refer to that the repurposed model might not continue to produce the same personal or even sensitive information as output. To stay with the example of a model that can predict the prevalence of hepatitis B: If that model is repurposed in a system that assesses job applicants, operators might do this in such a way that the 'prevalence of hepatitis B' is never explicitly evaluated or stored in an internal variable of the computing system. The model data of the original model might just get factored into a larger model that outputs yes/no decisions regarding whether an applicant should be invited to a job interview. It is therefore hard to prove for the data subjects in step 3, and easy for the creators and operators of the system to conceal, that the original hepatitis B prediction model was reused in creating the hiring decision model. This example showcases the importance of a regulatory approach that limits the ways in which model data that emerges from processing step 2 may be processed.

Purpose limitation for models implies that the creation and use of models are only permitted if the purpose for which this processing is done is named in advance and constitutes a valid purpose. In our proposal, we separate purpose limitation for models from an affected individual data subject. Given the collective and supra-individual risks involved, we do not consider it sufficient that only data subjects are able to control the processor's compliance with purpose limitation by means of individual rights. This is especially true as this is hardly ever the case in practice. Rather, a democratically legitimated institution should decide which purposes are desirable given the risks posed by the specific model. Thus, our goal is to escape the individualism that shapes the legal structures currently governing data processing in steps 1 and 3. Instead, we formulate an *ex ante* regulation with collective interests as the yardstick by which valid purposes are determined. After all, the hope that power asymmetries resulting from big data practices could be effectively contained by *ex post* regulation has not been fulfilled.[62] We therefore see purpose limitation for models as a tool to protect individual rights and interests of society different from creating new rights for data subjects. In the context of data practices that exploit the data of millions and could potentially impact everyone, placing the burden of responsibility on individuals appears misguided. Instead, we need to enable democratic participation of different stakeholders and empower the political collective to decide on the desired purposes of AI models.

2. Purpose Limitation for Models as Risk Prevention

Our starting point is the normative structure of risk prevention law, which deals in a preventive way with risks to individual rights and collective interests and societies. In environmental protection law, for example, there is ample evidence for the need to control and limit risks that originate from the actions of large and powerful actors using risk prevention legislation. Concerning generative and predictive machine learning models, we are facing an equivalent situation: first, in terms of content, there are many indications that predictive models harbour individual and societal risks. Due to the way the technology works, its enabling structure and its effects are collective – they affect a large number of individuals and unfold considerable spillover and leverage effects.[63] Second, these global technologies are difficult to address through individual-based legal systems and national enforcement mechanisms. Third, market mechanisms are less effective in digital markets than in analogue environments. This effect is due to a number of reasons: actors competing for new developments encourages predatory decisions. Furthermore, the market structures of digital markets are influenced

62 Tal Z Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2016) 47 Seton Hall Law Revie 995, 1011.

63 Omri Ben-Shahar, 'Data Pollution' (2019) 11 Journal of Legal Analysis 104, 105.

not only by the behaviour of actors but also by economies of scale or scope, network effects, switching costs, asymmetric and limited information, and consumer behavioural biases.[64]

There is no test of necessity for models based on aggressive data extraction from countless individuals. The benefit of targeted advertising over non-personalised advertising is unproven, yet it is offset by mass data breaches.[65] We argue for the adoption of the theoretical foundations of the principle of proportionality[66] from risk prevention law for the regulation of AI. Here, the normative structure of the test probing whether a means achieves its purpose in relation to the materiality of the restriction includes the protection of individual, collective, and political interests on the side of achieving the purpose. Simultaneously, informational asymmetries limit the benefits and interests of actors from the outset, mitigating the severity of regulation. In other words, the more people and critical legal interests are affected, the more regulation is justified. The more regulation is justified, the higher the level of democratic legitimacy should be at the level of structural and concrete decisions. In addition to the normative theoretical framework, the factual starting proposition is crucial: The societal benefits are still speculative in most areas, but the harms are empirically proven.

In this context, we understand purpose limitation for models as an instrument to preventively counter risks from specific AI models. Purpose limitation for societally risky models of certain powerful actors or in high-risk contexts of use would help redress power asymmetries and democratise the relevant context of use of AI. By applying the principle of proportionality from risk prevention law, parameters can be identified that point to a selection of high-risk models. Contrary to the proposal of the AIA,[67] we suggest to focus not on the intended use but on the position of the actors, the dissemination, the potentially affected legal interests and, above all, on the collective effects of these models.[68] The more likely it is that a large number of individuals or entire societies will be affected, the more a purpose limitation is justified. This risk may also arise because the actors involved are particularly powerful, have access to extensive databases, and exercise power similar to that of a state without being subject to fundamental rights due to their status as private actors. Data processing purposes should not be defined by the data-powerful actors themselves but by democratically legitimised specifications.

3. Criticism of Purpose Limitation in the Context of Big Data

Many voices have argued that big data and the purpose limitation principle are incompatible, especially in the context of the GDPR.[69] If the crucial point is that the GDPR should enable big data practices and promote the free flow of data (cf Article 1 (2) GDPR), this is convincing. To avoid this impasse, our regulatory proposal does not approach the problem from the big data side, which is the collection and processing of training data as described in processing step 1 (see section II.1). Rather, it seeks to regulate the processing of an entirely new kind of data that emerges only in processing step 2 (see section II.1). The trained model is often at the core of informational power asymmetry. To balance this power asymmetry, we apply purpose limitation only to the model data, leaving the general realm of big data practices untouched.

64 Johann Laux, Sandra Wachter and Brent Mittelstadt, 'Taming the Few: Platform Regulation, Independent Audits, and the Risks of Capture Created by the DMA and DSA' (2021) 43 Computer Law & Security Review 105613.

65 Veronica Marotta, Vibhanshu Abhishek and Alessandro Acquisti, 'Online Tracking and Publishers' Revenues: An Empirical Analysis', (Workshop on the Economics of Information Security, 2019).

66 For this see: Eric Engle, 'The History of the General Principle of Proportionality: An Overview' (2012) 10 Dartmouth Law Journal 1; David Duarte and Silva Sampaio (eds), *Proportionality in Law: An Analytical Perspective* (Springer Berlin Heidelberg 2018). In the context of EU law; Tor-Inge Harbo, 'The Function of the Proportionality Principle in EU Law' (2010) 16 European Law Journal 158. From a constitutional perspective: Matthias Klatt and Moritz Meister, *The Constitutional Structure of Proportionality* (OUP Oxford 2012). On the philosophical foundations: Marius Andreescu and Andra Puran, 'The Philosophical Basis of the Principle of Proportionality' [2022] Challenges of the Knowledge Society 188.

67 (n 58).

68 In regard to the proposed AIA, see also III.4. A comprehensive analysis why the proposed AIA does not sufficiently address the risks of unaccounted secondary use of trained models is currently under review in a separate publication, preprint: Rainer Mühlhoff and Hannah Ruschemeier, 'Updating Purpose Limitation for AI: A Normative Approach from Law and Philosophy' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4711621>.

69 Hildebrandt (n 42); Lokke Moerel and Corien Prins, 'Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things' (25 May 2016) <https://papers.ssrn.com/abstract=2784123>; Viktor Mayer-Schönberger and Yann Padova, 'Regime Change? Enabling Big Data through Europe's New Data Protection Regulation' (2016) 17 Science and Technology Law Review 315; Zarsky (n 62).

With this clarification in mind, the criticism that has been raised does not apply to our proposal. See, for instance Hildebrandt, who contends:

> [I]f Big Data is of interest because it generates patterns we could not have foreseen and thus enables usage that could not be predicted, then purpose binding is presumptuous and starts from the wrong premise. We do not know in advance what use is made possible, and to find out we must first mine the data [...]. The value of Big Data can only be set free if we admit the novelty of the inferred knowledge and rethink purpose binding in line with the innovative potential of its outcomes.[70]

This argument actually supports our proposal to shift the regulatory point of intervention from big data to model data and its context of use. Open purpose data mining can be seen as an exploratory process not immediately tied to an application. If the data mining involves building models, our purpose limitation procedure would require including something like 'foundational research' as a valid purpose. This procedure does *not* apply any restrictions to the research and therefore enables researchers to 'mine' the potential of big data. At the same time, however, our proposal ensures that if applicable models emerge during this foundational research, they cannot immediately be put to practical use. This is a feature, not a limitation, of our regulatory proposal. If resulting models are intended to be used in application domains, reaccreditation of the new purpose is required.

Another version of the critique uses Nissebaum's philosophical framework of 'privacy as contextual integrity' as a way out of the perceived impasse surrounding purpose limitation for big data.[71] For instance, Hahn elaborates that contextual integrity, although helping to balance various 'informational norms' relevant to the context from which big data is collected, could potentially be violated as a result of open purpose data processing.[72] Hahn contends that this approach allows for a more nuanced ethical consideration of the validity of big data practices where the principle of purpose limitation fails with respect to large data companies (cf pages 41–42). As she argues:

> Therefore, the argument is advanced that the contextual integrity framework can be used as a starting point to warrant the stricter enforcement of the Purpose Limitation principle with regards to Data Power companies in particular. It is proposed that the framework be used to evaluate the consequences of failing to respect Purpose Specification, in order to show that these violate the expectations of the data subject.[73]

While it may be true that contextual integrity allows for a more nuanced analysis of the privacy of the data subjects in the training data, a similar objection as before applies: in this article, we are not interested in regulating processing step 1 (see II.1), ie the collection and processing of big data as training data. Rather, we propose a purpose limitation concerning the model data emerging from step 2 (see II.1). This then regulates how this data may be processed with respect to potential applications in step 3 (see II.1). Since model data must be assumed to be anonymous and highly aggregated data,[74] there is no data subject to which Hahn's argument could apply. In particular, it is not immediately plausible what kind of moral wrong is done to the data subjects in the training data if the model data is utilised in harmful ways. Instead, we are addressing the need for *preventative* protection of anyone from the potential harms resulting from applications of the model data.

4. Problems with Regulating Purposes

Limiting purposes is a demanding regulatory goal. Objectives can be formulated at different levels of abstraction and from various perspectives. Therefore, it is essential to determine how and by whom the purposes of existing and prospective models are to be defined. Objective third parties, stakeholders, or the model's users can all define purposes for an intended use. The current regulatory approach of the AIA is to regulate risk according to the intended use and extent of use by the system provider, currently

70 Hildebrandt (n 42).

71 Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford University Press 2009); Hildebrandt (n 42), 37 et seq.

72 Hahn (n 45).

73 Ibid, 43.

74 Representing 'generalised knowledge,' Michele Loi and Markus Christen, 'Two Concepts of Group Privacy' (2020) 33 Philosophy & Technology 207.

defined by Articles 6 (1, 2), 7 (2) (a) AIA.[75] We are critical of this for several reasons:

In its regulatory approach, the AIA is designed as a product liability regime rather than a primary instrument for safeguarding fundamental rights or addressing societal risks.[76] This is also demonstrated by the fact that the AIA considers the trustworthiness of a system decisive for accepting the associated risks. However, the societal risks we seek to address in this paper cannot be resolved by the yardstick of trustworthiness. The problem posed by power asymmetries is not reflected in the framework of the AIA. Its risk classification scheme does not consider the position of actors within power relations with the exception of a few sectoral regulations for small and medium enterprises. The contexts of use listed in Annex III[77] do, in some cases, coincide with the problematic purposes for which models are used. We maintain that here too it is the particular form of exercise of power that should become the objective of improved regulation.

Classifying risk based on intended use is closely linked to standards that standardisation organisations have yet to define for AI systems (cf Article 42 (1)). This approach does not adequately reflect the complex power structures involved in the deployment of AI technology because reliance on voluntary standards and self-governance is 'disregarding power-related considerations'.[78] Further, there is no risk assessment by an independent entity. In this situation, as has already been criticised,[79] the central actors referred to by the AIA are not the providers or users but the European standardisation organisations European Committee for Standardisation (CEN) and European Committee for Electronic Standardisation (CENELEC). These bodies are in charge of developing harmonised standards (cf Articles 40 *et seq)*. As a result, the AIA lacks the substantive legal requirements and the socio-technical context of systems: When should discrimination be forbidden? When is human oversight meaningful? What ethical standards should apply to systems?[80] As a deliberate political choice, the AIA thus outsources the core ethical and regulatory questions to private organisations. This is problematic because they lack sufficient stakeholder participation and democratic legitimation.[81]

The example of the recent provisions for general purpose AI shows that an orientation towards *ex ante* defined purposes of use by providers reaches its limits with systems that can be used for various purposes. The Council proposal for the AIA[82] provided that general purpose AI systems are considered high risk if they can be used in the sense of Articles 6, 4 (b) (1) AIA. This is circular, however, as Article 6 AIA is based on the intended use and the extent of the use. Consequently, the providers are not held responsible and the risk is shifted to the users. Article 4c paragraph 1 provides for the requirements of Article 4b to not apply if the provider has excluded high-risk uses in the instructions for use or accompanying documents.[83] Following this provision means that suppliers can absolve themselves of the responsibility through a formal exclusion clause. This leaves no safeguards in place to ensure users do not utilise the system in a high-risk way, even if it is challenging to exclude all potential high-risk scenarios.[84] Similar to the GDPR, only well-funded, globally active players have the resources to draft appropriately worded exclusion clauses to comply with the requirements of the AIA.

As an alternative approach, our proposal of purpose limitation for models includes listing permissible purposes, relevant actors, and acceptable contexts

75 Similar in the Commission's proposal and the European Parliament amendments (n 58).

76 Cf Marco Almada and Nicolas Petit, 'The EU AI Act: A Medley of Product Safety and Fundamental Rights?' (European University Institute 2023) Working Paper <https://cadmus.eui.eu/handle/1814/75982>; art 6(1) AIA.

77 In all versions (n 58).

78 Maciej Kuziemski and Gianluca Misuraca, 'AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings' (2020) 44 Telecommunications Policy 101976.

79 Johann Laux, Sandra Wachter and Brent Mittelstadt, 'Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act' (20 February 2023) <https://papers.ssrn.com/abstract=4365079>; Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22 Computer Law Review International 97.

80 Cf Hannah Ruschemeier and Rainer Mühlhoff, 'Daten, Werte Und Der AI Act: Warum Wir Mehr Ethik Für Bessere KI-Regulierung Brauchen' [2023] Verfassungsblog <https://verfassungsblog.de/daten-werte-und-der-ai-act/>.

81 Veale and Borgesius (n 79); Nathalie A Smuha et al, 'How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act' (5 August 2021) <https://papers.ssrn.com/abstract=3899991> accessed 14 July 2023.

82 2021/0106(COD), 15698/22 [2022].

83 This no longer appears to be the case in the agreement on the draft, but this document was not yet officially available at the time this paper was submitted.

84 Hacker, Engel and Mauer (n 3).

for using AI models that are to be defined in democratic processes. Rather than classifying the risk associated with a system according to individual and intended purposes of use as declared by the providers, our approach resolves the conflicts between intended purposes and generative or general purpose AI by identifying societally beneficial purposes that align with the principle of proportionality.

IV. Conclusion and Outlook

AI technology implies risks and benefits that affect billions of people worldwide. The general call for effective regulation of this technology arises from the legitimate concern that in the *status quo,* private companies, driven solely by their economic interests, have unchecked control over the implementation of this technology. In this paper, we introduced the concept of purpose limitation for models as a regulatory approach to one severe risk associated with machine learning technology: the risk of unaccounted and potentially harmful secondary use of trained models. A legal approach to mitigating this risk should focus on robust state and administrative structures, namely effective oversight and enforcement.

On the way to a legal implementation of purpose limitation for models, a lot of work is yet to be done. Our proposal seeks to establish a democratic discourse about the legitimate purpose of AI. Any implementation should involve diverse stakeholders in establishing categories and norms that govern purpose limitation for models. Whilst the detailed elaboration of a positive list and the underlying ethical principles is the subject of a separate work to be conducted with participatory methodology, our aim in this paper was to introduce the conceptual idea of a regulatory approach to govern trained models concerning their permissible uses. In the following, we will provide some foundational remarks on a legal implementation of purpose limitation for models.

Relying solely on individual data subjects' rights is insufficient to address the power asymmetries interwoven with AI technology and specifically the risks that result from unaccounted reuse of trained models. Therefore, purpose limitation should not only be enforced through individual-protecting measures like the ones included in the GDPR but should be anchored at a systemic level. This approach enables a regulatory regime that imposes special obligations on data-powerful actors who structurally violate data protection and privacy, even outside of competition law.[85] Actors that fundamentally and structurally threaten and undermine accountability structures should be subject to special requirements. Implementing this systemic approach would extend the privileges granted for (non-commercial) statistical purposes and scientific research (cf Articles 5 (1) (b), 89 GDPR) beyond the confines of the GDPR.

Proposing a definition of the permissible purposes for societally risky AI models might seem like a radical idea, especially if this definition aims to be more precise than the existing data protection principle of purpose limitation. However, a similar approach is already adopted in the proposed Regulation on the European Health Data Space (EHDS).[86] The EHDS proposal clearly defines the purposes for electronic health data processing for secondary use in Article 34 and additionally names prohibited secondary data use in Article 35. The 'positive list,' for example, includes purposes which are activities in the public interest. These comprise public health surveillance and protection against cross-border threats (a), supporting public sector bodies (b), producing statistics (d), and education or teaching (e). On the one hand, Article 34 explicitly allows for development and innovation activities for the quality and safety of health care (f), the training, testing, and evaluating of algorithms, including in medical devices, AI systems, and digital health applications that contribute to public health or social security (g), and providing personalised healthcare consisting of assessing, maintaining, or restoring the state of health of natural persons based on the health data of other natural persons (h). On the other hand, Article 35 excludes purposes such as taking decisions concerning natural persons or groups of natural persons that would exclude them from the benefit of an insurance contract or modify their contributions and insurance

85 For the interaction between data protection and competition law, see: ECJ C-252/21; Philipp Hacker, 'Manipulation by Algorithms. Exploring the Triangle of Unfair Commercial Practice, Data Protection, and Privacy Law' [2021] European Law Journal 1; Orla Lynskey and Francisco Costa-Cabral, 'Family Ties: The Intersection between Data Protection and Competition in EU Law' (2017) 54 Common Market Law Review; Hannah Ruschemeier, 'Competition Law as a Powerful Tool for Effective Enforcement of the GDPR' (*Verfassungsblog*, 7 July 2023) <https://verfassungsblog.de/competition-law-as-a-powerful-tool-for-effective-enforcement-of-the-gdpr/>.

86 Com2022/197-final.

premiums, Article 35 (b), as well as advertising or marketing activities towards health professionals, organisations in health or natural persons, Article 35 (c).

Documenting models that pose a particular risk to society is a necessary first step for compliance and quality control. Following this, an oversight body should be set up at European Union level to establish guidelines for a limited period of time for purpose identification, legal implementation, and enforcement. The AIA envisions something similar in the form of regulatory sandboxes at the Member State level (cf Articles 53–55a).[87] This way, regulators can test innovative AI applications for a limited period. However, such approaches should not only be concerned about fostering innovation but also about regulatory learning. It is crucial to address the contexts of AI use by developing procedures for purpose limitation in a temporary framework that can successfully be introduced into the political and legislative process. As a result, both sector-specific regulation[88] and the creation of a new supervisory authority[89] may prove to be viable options as well as strengthening collective redress mechanisms.[90] The monitoring of compliance with the stated purpose is a permanent obligation. Its implementation can be borrowed from the Digital Services Act (DSA): trusted flaggers, transparency and reporting obligations, and the monitoring of systemic risks.

In conclusion, a positive list of permissible uses of trained models needs to be developed. Combined with our proposed regulatory approach of a purpose limitation for models, this enables balancing of providers' interests and individual, collective, and societal risks. In our view, this would be a meaningful step towards regulating the use of AI models. This approach would consider the actors' position within the informational power asymmetries arising in the context of AI and the global impact of their applications. In the interaction between global Big Tech companies and users, it can no longer be assumed that private actors are facing each other on equal terms. Therefore, societal impact should play a much more significant role in risk classification than it has in the past. The methodological tool of proportionality assessment to identify risks to be addressed by regulation should be based on ethical considerations and stakeholder participation as we will elaborate in future research.

87 In all versions (n 58).

88 Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 UCLA Law Review 1701.

89 Andrew Tutt, 'An FDA for Algorithms' (2017) 69 Administrative Law Review 83.

90 Hannah Ruschemeier, 'Kollektiver Rechtsschutz und strategische Prozessführung gegen Digitalkonzerne' (2021) 24 MMR 942.