LAW, INNOVATION AND TECHNOLOGY
https://doi.org/10.1080/17579961.2024.2313794

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Predictive analytics and the collective dimensions of data protection

Rainer Mühlhoff[a] and Hannah Ruschemeier[b]

[a]Institute for Philosophy & Institute for Cognitive Science, University of Osnabrück, Osnabrück, Germany; [b]Department of Law, University of Hagen, Hagen, Germany

ABSTRACT
This paper takes an interdisciplinary approach, combining legal studies, ethics and technical insights to shed light on the complex issues surrounding the regulation of predictive analytics. We argue that the individualised concept of regulation, shaped by the dogma of fundamental rights, is unable to adequately capture the implications of predictive analytics. We show that predictive analytics is a problem of collective privacy and informational power asymmetries, and conceptualise the form of data power at work in predictive analytics as 'prediction power'. The unregulated prediction power of certain actors poses societal risks, especially if this form of informational power asymmetry is not normatively represented. The article analyses this legal lacuna in the light of recent case law of the European Court of Justice and new legislation at the EU level. To address these challenges, we develop the concept of 'predictive privacy' as a protected good based on collective interests.

ARTICLE HISTORY Received 11 July 2023; Accepted 23 October 2023

KEYWORDS big data; ethics; DSA; GDPR; collective data protection; privacy

1. Introduction[1]

Debates about artificial intelligence (AI), big data, and the societal influence of large information and communication technology (ICT) corporations remain constrained by the boundaries of their own disciplines, especially those discussions approached from a legal perspective.[2] As we will show,

CONTACT Hannah Ruschemeier ✉ hannah.ruschemeier@fernuni-hagen.de

[1] An earlier, shorter and preliminary version of this research was published in German without covering current developments in EU legislation and case law under: Rainer Mühlhoff and Hannah Ruschemeier, 'Predictive Analytics und DSGVO: Ethische und rechtliche Implikationen.', in Hans-Christian Gräfe and Telemedicus e.V. (eds.), *Telemedicus – Recht Der Informationsgesellschaft, Tagungsband Zur Sommerkonferenz 2022* (Frankfurt am Main: Deutscher Fachverlag 2022), 38–67.

[2] Wolfgang Hoffmann-Riem, 'Artificial Intelligence as a Challenge for Law and Regulation' in Thomas Wischmeyer Timo Rademacher (eds), *Regulating Artificial Intelligence* (Springer Nature 2020), Ralf

one area where interdisciplinarity between law and philosophy promises to be particularly fruitful is the critical analysis of predictive analytics — that is, the use of AI and statistical computing techniques to derive predictions about individuals or groups.

Predictive analytics roughly refers to the practice of using machine learning models for prediction purposes and is currently one of the most important applications of big data and machine learning technology. Predictive analytics constitutes a novel challenge to privacy and data protection legislation as it is often used in the context of algorithmic scoring and automated decision making to infer unknown personal information about individuals, or to categorise individuals according to such estimates in order to treat them differently. In this paper we will take an interdisciplinary approach involving legal studies, ethics and technical insight to offer a fresh take on the challenges of regulating predictive analytics.

Predictive models are typically trained on large sets of training data from which a machine learning procedure can 'learn' correlations or 'patterns' in the data. Trained models can then be used to automatically classify individuals or cases according to criteria such as business risk, health risk, psychological traits, substance abuse, creditworthiness, consumer interests, political views, religious affiliations, sexual identity, etc. These systems allow the operating party, which is typically a large data or platform company, to make a predictive assessment on a case-by-case (individual) basis of unknown (personal) information — unknown either because they are unknown to the assessing party as the actual information is hard to obtain (e.g., health information, sexual identity) or because the predicted information is about future events (e.g., business risks, credit default, purchasing behaviour) that may not even be known to the person concerned.[3] Predictive systems are commonly used in targeted advertising, differential pricing, credit scoring, insurance risk assessment, automated hiring systems, but also in predictive policing or determining potential risk of recidivism.[4]

Our initial thesis starts from the concern that predictive analytics technology has potentially significant societal impact that scales with the wide proliferation of this technology. One of the risks that has often been mentioned is unfair biases in these models — that is the systematic and unfair discrimination against certain individuals or groups.[5] Our article complements the

Poscher, 'Artificial Intelligence and the Right to Data Protection' in Silja Voeneky et al Wolfram Burgard (eds), *The Cambridge Handbook of Responsible Artificial Intelligence* (Cambridge University Press 2022).

[3] Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer, 2008).

[4] Danielle K Citron and Frank A Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) *Washington University Law Review* 2.

[5] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' [2018] Conference on Fairness, Accountability and Transparency; Cathy O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy* (Penguin

debate on bias by shifting attention to another and even more fundamental risk arising from predictive analytics, which is a specific invasion of privacy: Predicting personal information about arbitrary individuals or groups can, we argue, constitute an invasion of privacy as much as the unauthorised use of factual information can (by factual information here we mean information that has been disclosed by the data subject, as opposed to predicted information).[6] This leads to a transformation of social relations, creating dangers for democracy and social (group) harmonisation, as well as ubiquitous commercialisation. As we will specifically discuss in this article, the privacy risks of predictive analytics already arise from *creating* predictive models (as opposed to the later step of applying the model to a concrete case).[7]

Hence, predictive analytics' mode of operation leads to direct legal implications: infringements of privacy,[8] discrimination,[9] and the exploitation of power asymmetries that defeat the individual and collective realisation of rights.[10] However, the current legal framework, especially as represented by the EU's General Data Protection Regulation (GDPR), and also the Digital Services Act (DSA), falls short in terms of implementing sufficient control over the societal risks of predictive analytics. Automated predictions can harm both individuals and entire societies. At the same time, predictive analytics relies on a structure of collective causation, insofar as prediction systems exploit the collective data provided by many data subjects upon using digital services and apply it to targets that may even come from outside that group of data subjects.[11] As predictive analytics is only feasible for those actors who hold large amounts of user data, we analyse its exercise as a specific manifestation of informational power asymmetry in the form of

Books, London 2017); Safiya U Noble, *Algorithms of oppression: How search engines reinforce racism* (New York University Press 2018).

[6] Rainer Mühlhoff, 'Predictive privacy: towards an applied ethics of data analytics' (2021) *Ethics Inf Technol* 675.

[7] On AI models as an object of regulation: Rainer Mühlhoff and Hannah Ruschemeier, 'Democratising AI via Purpose Limitation for Models', preprint (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4599869; 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), *KI und Robotik in Der Medizin – Interdisziplinäre Fragen* (Nomos 2024).

[8] Adrian Kuenzler, 'What competition law can do for data privacy (and vice versa)' [2022] 47 *Computer Law & Security Review*. On Big Data in general: Ira S Rubinstein, 'Big Data: The End of Privacy or a New Beginning?' (2013) 3 *International Data Privacy Law* 74.

[9] Sandra Wachter, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law (2023) 97 *Tulane Law Review* 149; Sandra Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (2019) *Berkeley Technology Law Journal* 1; Raphaële Xenidis and Linda Senden, 'EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination' in Ulf Bernitz, Sybe de Vries (eds), *General Principles of EU Law and the EU Digital Order* (Wolters Kluwer 2020).

[10] Omer Tene and Jules Polonetsky, 'Big Data for All: Privacy and User Control in the Age of Analytics' [2013] 11 *Northwestern Journal of Technology and Intellectual Pro*perty 239.

[11] Mühlhoff and Ruschemeier (n 1); Mühlhoff (n 6).

prediction power. We contend that instead of addressing these informational power asymmetries, the GDPR reproduces them.[12]

Millions in fines[13] and an increasing number of lawsuits[14] against digital companies do not address the fundamental problem of predictive modelling as a core business model. One reason is that large technology companies have been very successful in evading various forms of regulation, refusing to acknowledge the systemic and collective risks of their business models and deliberately exploiting power asymmetries through their own norm-setting and ubiquitous distribution. Additionally, the architecture of European legal protection poses a systemic challenge to the normative capture of collective causal structures and the impact of predictive analytics which leverages the data of some individuals to make predictions about others. This is because the legal system is designed to enforce the rights of individual data subjects,[15] not regulate the collective impact of this technology.[16]

The jurisprudence of the European Court of Justice (ECJ)[17] has also only indirectly addressed the challenges of predictive analytics in terms of collective implications, but the current decision on the data pratices of the company Meta, on the other hand, is a milestone against collective data exploitation.[18] The paper analyses its impact. We argue that a new understanding of privacy is needed to identify these regulatory gaps and to develop proposals for solutions. We contend that these normative considerations are inextricably linked to the technical foundations and ethical theorising of privacy as a subject of protection, and that ethical and legal perspectives are complementary rather than distinct.[19]

This article starts with a short outline of the technical prerequisites of predictive analytics and the philosophical and ethical concept of prediction

[12] See further Hannah Ruschemeier, ‘Data Brokers and European Digital Legislation’ (2023) 9 *European Data Protection Law Review* 27.

[13] E.g., the decision of the Irish Data Protection Commission: IN-21-4 in the matter of Meta Platforms Ireland Ltd., Decision of the Data Protection Commission made pursuant to Section 111 of the Data Protection Act 2018 and Article 60 of the General Data Protection Regulation, https://www.dataprotection.ie/sites/default/files/uploads/2022-12/Final%20Decision_IN-21-4-2_Redacted.pdf.

[14] E.g., www.justice.gov/opa/pr/justice-department-sues-google-monopolizing-digital-advertising-technologies.

[15] Predictive analytics does not rely on the deliberate disclosure of information about third parties by individuals. However, Johannes Eichenhofer, *e-Privacy: Theorie und Dogmatik eines europäischen Privatheitsschutzes im Internet-Zeitalter* (Jus Publicum, Mohr Siebeck, Tübingen 2021) 152 appears to only assume a threat to privacy arising from deliberate disclosure of information via third parties.

[16] In this context of genetic data: Taner Kuru and Iñigo d M Beriain, ‘Your genetic data is my genetic data: Unveiling another enforcement issue of the GDPR’ [2022] 47 *Computer Law & Security Review*.

[17] ECJ C-184/20, Vyriausioji tarnybinės etikos komisija [2022] (ECLI:EU:C:2022:601); ECJ Case C–141/12 YS and Others [2014] (ECLI:EU:C:2014:2081) and ECJ Case C–372/12 M and S [2012] (ECLI:EU:C:2014:2081); ECJ Case C-434/16 Nowak [2017] (ECLI:EU:C:2017:994).

[18] ECJ-C252/21 *Meta vs. Bundeskartellamt*.

[19] On the relationship of law and ethics in the context of AI: Hannah Ruschemeier and Rainer Mühlhoff, Daten, Werte und der AI Act, https://verfassungsblog.de/daten-werte-und-der-ai-act/ (accessed 26 December 2023); Giovanni Sartor, ‘Artificial intelligence and human rights: Between law and ethics’ (2020) 27 *Maastricht Journal of European and Comparative Law* 705.

power, before analysing the legal implications. Subsequently, we explain why privacy, as a legally protected right, depends on insights from other disciplines and is therefore particularly open to development. We introduce our concept of predictive privacy and distinguish it from related proposals such as group privacy. We also discuss the view that neither the GDPR nor anonymisation techniques can be effectively positioned against the collective structures of predictive analytics. We clarify that the described challenges are not solved by the DSA and conclude that effective protection of privacy and personal data requires systemic risks to be better addressed in future legislation.

Overall, we offer an interdisciplinary understanding of data privacy and data protection law by pointing out the technical, ethical, and legal implications to cultivate a nuanced debate in the light of the new European legislation.

2. Technical basics of predictive analytics

2.1. Concept and technical procedure

By the term 'predictive analytics' we refer to certain applications of 'artificial intelligence' (AI), data analysis, and computational statistics techniques that use available data to build predictive models. By a predictive model we mean an algorithmic routine used to estimate unknown or future information (target information) about an individual or case from auxiliary data. Such predictions may concern the behaviour of people, events, or be used to classify people into similarity groups. In this context, a predictive model typically receives as input the information known about an individual or case to be assessed (hereafter referred to as 'auxiliary data', e.g., tracking data or social media usage data about a user) and returns as output an estimate of the modelled target variable (e.g., sexual identity of the user).[20]

Predictive models are typically trained or calibrated using large amounts of training data. Such training data sets consist of data pairs that combine auxiliary data and target information over a large number of known cases. Learning methods that train a model using this kind of training data are also called 'supervised' learning methods because they learn from examples for which the target variable to be modelled is already known.

2.2. Examples of application

Generally speaking, predictive modelling is of interest when and where difficult-to-access information about any user is estimated on the basis of

[20] Mühlhoff (n 6); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (Public Affairs 2020).

readily available data. For example, medical researchers at the University of Pennsylvania have shown that usage data from social media platforms can be used to predict whether a user suffers from diseases such as depression, psychosis, diabetes, or high blood pressure.[21] A well-known study by *Konsinski* et al. determined that a Facebook user's likes can be used to predict 'a range of highly sensitive personal attributes' about that user, 'including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, addictive behaviour, parental separation, age, and gender.'[22] Other important application areas[23] of predictive analytics are targeted advertising[24] and credit scoring.[25]

Beyond the apparent implication that predictive analytics enables violations of the target subjects' informational self-determination, of which most users are unaware, the *secondary* use of such predictions is particularly relevant. Such predictive analyses are of great interest, for example, to insurance companies or in the context of recruitment procedures, because they allow individual risk measurement. Insurance companies can also use predictive models to 'nudge' their customers individually through, for example, the use of targeted discount programmes.[26] Such methods can become a significant data protection risk if the object of targeting is, for example, advertising for medical products or drugs, in which case the predictive model trained in the method ultimately enables the prediction of diseases and thus sensitive medical information for any platform user.[27]

2.3. Two processing steps: technical implications

In discussing the ethical and legal implications of predictive analytics, it is helpful to bear in mind that applications of this technology involve *two* steps of data processing that must be considered separately.

We call the first step the *training of a predictive model*. Here, suitable data from a large number of individuals, users, or incidents are collected in order

[21] Raina M Merchant et al, 'Evaluating the Predictability of Medical Conditions from Social Media Posts' 14 1932-6203 e0215476.

[22] Michal Kosinski, David Stillwell and Thore Graepel, 'Private traits and attributes are predictable from digital records of human behavior' [2013] PNAS. See also the examples in Zuboff (n 20) 129 et seq.

[23] For a comprehensive overview see Danielle Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89 *Washington Law Review* 1.

[24] E.g., Johann Laux, Sandra Wachter and Brent Mittelstadt, 'Neutralizing Online Behavioural Advertising: Algorithmic Targeting with Market Power as an Unfair Commercial Practice' (2021) 58 *Common Market Law Review* 719.

[25] Mikella Hurley and Julius Adebayo, 'Credit Scoring in the Era of Big Data' (2017) 18 *Yale Journal of Law and Technology* 5.

[26] Christian Roth et al., 'Are Sensor-Based Business Models a Threat to Privacy? The Case of Pay-How-You-Drive Insurance Models' in Stefanos Gritzalis et al. Ismail Khalil (eds), *Trust, Privacy and Security in Digital Business* (Springer International Publishing 2020).

[27] As argued by Rainer Mühlhoff and Theresa Willem, 'Social media advertising for clinical studies: Ethical and data protection implications of online targeting' (2023) 10 *Big Data & Society* 1–15.

to train a predictive model in the sense of (2.1), using suitable (machine learning) methods. The training data can also be anonymised data as long as it consists of a pair of 'auxiliary data' and 'target information' for each individual or incident covered in the training data set (e.g., users' Facebook likes as auxiliary data, and their explicitly provided information about their sexual identity as target information; actual identifying information about the users can be deleted). The predictive model produced in this first processing step is fundamentally not personal data; it is a calibrated algorithmic routine to estimate the target variable about *any* individual or incident of which the auxiliary data is, or becomes, available at any time *in the future*.

We refer to the second data processing step involved in the application of predictive analytics as the *inference step*. Here, an existing predictive model is applied to a specific individual or incident in order to estimate the target information about that specific case. As a rule, the target individual in this step is known and identifiable (even if they appear as a pseudonymous user on a platform). Auxiliary data collected about the target individual (e.g., tracking data, Facebook likes) is used as case-based input data for this step; the predictive model is applied to this input data to predict target information about the individual. In general, personal and sometimes sensitive information about an identifiable individual is obtained in this second processing step.

The second processing step does not need to follow immediately upon the first and does not need to be conducted by the same entity. In fact, models could be trained by data companies (first step) and then be stored for arbitrary time, distributed to third parties, pushed to end-user devices etc., before an actual inference (second step) is computed from these models.

3. Privacy challenges arising from predictive analytics

3.1. Prediction power as a current manifestation of data power

The privacy threat of predictive analytics arises already in the *first* data processing step, before (and independent from) any inferences that are actually computed from the model.[28] As a result of the training step, while there has not yet been a specific privacy violation for a specific individual, there is an actor with a predictive model at their disposal who has the *ability* to estimate certain information about *any* individual on the basis of auxiliary data.[29] Possession of such a model presents an invasion of anyone's privacy *in*

[28] Rainer Mühlhoff, 'Predictive privacy: Collective data protection in the context of artificial intelligence and big data' (2023) 10 *Big Data & Society*.

[29] We therefore argue that the possession of certain models must be regulated: Rainer Mühlhoff and Hannah Ruschemeier, Democratising AI via Purpose Limitation for Models, preprint (2023) https://doi.org/10.2139/ssrn.4599869 .

potentia, that is, an invasion that potentially affects and ‘threatens’ broad strata of society. Therefore, what we call *prediction power* emerges in the first processing step,[30] posing a not yet actualised capacity to perform certain ethically and legally relevant actions.

We argue that the instruments of the GDPR are not sufficient to effectively address the societal and individual risks posed by prediction power, which mostly accumulate in the private sector but which may also arise in public data-processing organisations. A new protection concept is therefore necessary to legally codify and effectively regulate the prediction power of individual actors that arises through data accumulation.

To reach this goal, the present project adopts a concept of privacy and data protection that is not limited to the protection of individual privacy, but which serves to balance informational power asymmetries between society and data-processing organisations. ‘Privacy is indeed about power’, as *Ari Ezra Waldman* puts it.[31] Orienting data protection towards balancing data power has long been debated and occasionally implemented;[32] the new contribution of our approach lies in pointing out that prediction power is the most current manifestation of data power.

3.2. Combining ethical and legal concepts of privacy

The concepts of privacy and data protection are related but not identical. In legal terms, privacy is one of the protected goods of data protection law. In ethics, on the other hand, privacy is understood as a value and theorised in a variety of different historical and social contexts.[33] In contrast to this comprehensive debate, the protection of privacy in data protection law is specifically limited to the processing of personal data.[34] At the same time, data

[30] Rainer Mühlhoff, ‘Prädiktive Privatheit: Kollektiver Datenschutz im Kontext von Big Data und KI’ in Michael Friedewald Alexander Roßnagel (eds), *Künstliche Intelligenz, Demokratie und Privatheit* (Nomos, Baden-Baden 2022).

[31] Interview with Ari Ezra Waldmann in Sarah Kardesler, ‘Why are we not building a queer movement around privacy?’ (*PinG*, 23 December 2021) https://www.pingdigital.de/blog/2021/12/23/why-are-we-not-building-a-queer-movement-around-privacy/2230.

[32] Brent Mittelstadt, ‘From Individual to Group Privacy in Big Data Analytics’ (2017) 30 *Philosophy & Technology* 475; Angelina Fisher and Thomas Streinz, ‘Confronting Data Inequality’ (2021) 60 *Columbia Journal of Transnational Law* 829; Ari E Waldman, *Industry unbound: The inside story of privacy, data, and corporate power* (Cambridge University Press 2021); Daniel J Solove, ‘Privacy and Power: Computer Databases and Metaphors for Information Privacy’ (2001) 53 *Stanford Law Review* 1393; Simone van der Hof and Corien Prins, ‘Personalisation and its Influence on Identities, Behaviour and Social Values’ in Mireille Hildebrandt Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008) 118.

[33] Priscilla M. Regan, ‘Privacy and the common good: revisited’ in Beate Roessler Dorota Mokrosinska (eds), *Social Dimensions of Privacy: Interdisciplinary Perspectives* (Cambridge University Press 2015); Herman T Tavani, ‘Philosophical Theories of Privacy: Implications for an adequate Online Privacy Policy’ [2007] 38 *Metaphilosophy* 1.

[34] David H Flaherty, ‘On the utility of constitutional rights to privacy and data protection’ (1990) 41 *Case Western Reserve Law Review* 831; Eichenhofer (n 15).

protection is not limited to the protection of privacy; this is reflected, for example, in the European Charter of Fundamental Rights, which understands the right to data protection in its scope of data processing as going beyond the protection of the right to privacy.

'Privacy' as it figures in law is not a factual or normatively predefined protected good that is typically wrapped up in a specific area of law. In contrast to institutional guarantees such as property, or factual predefinitions such as the protection of human life, privacy is a social construct.[35] The assessment of whether privacy as a value requires protection must necessarily draw on the assumptions of other disciplines, as it cannot be concluded legally.[36] Therefore, legally standardised forms of privacy protection are particularly open to development and new interdisciplinary insights such as the concept of predictive privacy.

3.3. Ethical and epistemic implications of predictions

In this paper we build on the 'predictive privacy' concept as introduced by *Mühlhoff*.[37] A violation of predictive privacy occurs when personal information about an individual or group is *predicted* — this could include information that the data subject did not explicitly disclose anywhere or does not even know themself (e.g., the prognosis of a disease). The fact that privacy could be violated by way of prediction, particularly predictions made on the basis of mass data collected from many *other* individuals, has so far received little attention in politics and the public discourse. For many, it is not part of the moral consciousness around data protection and privacy on the internet, and similarly has received little attention from ethicists and academic circles.

There is the particular problem of the 'prediction gap' that forms when algorithmic predictions are translated into (automated) decisions that lead to action.[38] While the output of a predictive model is generally a vector of possible values of the target variable weighted with individual probabilities (e.g., a model that predicts sexual identity from social media data might output something like '80% heterosexual', '15% homosexual' …), an automated decision based on such a prediction would imply picking the best

[35] Waldmann argues that Privacy Law participates in the social construction of new technologies that make surveillance easier Ari E Waldman, *Privacy as trust: Information privacy for an information age* (Cambridge University Press, 2018); on the interdisciplinary aspect: Christoph Gusy, 'Was schützt Privatheit? Und wie kann Recht sie schützen?' (2022) 70 *Jahrbuch des öffentlichen Rechts der Gegenwart. Neue Folge (JöR)* 415.

[36] On Privacy as a socially constructed value: Julie E Cohen, 'What Privacy Is For' (2013) 126 *Harvard Law Review* 1904; Nicholas Proferes, 'The Development of Privacy Norms', *Modern Socio-Technical Perspectives on Privacy* (Springer 2022) 81–82*;* Brian Shapiro and C.Richard Baker, 'Information technology and the social construction of information privacy' [2001] 20 *Journal of Accounting and Public Policy* 295.

[37] Mühlhoff (n 6); Mühlhoff (n 28).

[38] Mühlhoff (n 6).

match from this vector (e.g., the option with the maximum probability weight). The person is then automatically treated *as if* they already possess this property. This kind of reasoning implies the transformation of a statistical inference, which is always knowledge related to the population as a whole, into a prediction about an *individual case* (point prediction). This step is not covered by the logics of classical statistics as it corresponds to 'making a bet' about the individual.[39] This mechanism of betting on the individual and treating them as if they already manifest a certain (in reality, unknown) trait constitutes a limitation of individual autonomy, which is an ethical problem specific to this type of data processing.[40]

Another specific feature of the ethical problem of predictive privacy is its *collective causation*. Violations of predictive privacy through predictive analytics are only possible if many individuals disclose the data about themselves necessary to produce the models, combined with a lack of regulation preventing data-processing organisations from using that data to produce predictive models.

This problem of collective enabling is similar to the problem of individual greenhouse gas emissions in the context of climate change: here, too, individual emissions act as societal externalities, i.e., as costs to society that affect everyone and are not priced into the individual cost–benefit considerations of users.[41] However, in contrast to the question of whether one should use a high-emission or low-emission means of transport for the good of society, for example, the decision of whether one should use a certain data-based service is framed in the current societal discourse exclusively as an *individual* decision and cost–benefit consideration, while the potential collective effects in relation to data protection consequences are discursively ignored.

3.4. Predictive privacy as a legally protected interest – in conflict with data protection?

The fact that predictive models can only function by processing a large number of data from different persons is currently not reflected in the law, as seen in the GDPRs limitation to personal data. Inferred data, even from anonymised data, may allow conclusions to be drawn about highly personal characteristics. Limiting the GDPR to the processing of personal data therefore does not consider the collective element of modelling.

[39] See generally Justin Joque, *Revolutionary mathematics: Artificial intelligence, statistics and the logic of capitalism* (Verso 2022).

[40] Mühlhoff (n 6).

[41] Following these lines of thought, the concept of 'data pollution', for example, has been proposed, see Omri Ben-Shahar, 'Data Pollution' (2019) 11 *Journal of Legal Analysis* 104.

It is a dilemma of European fundamental rights[42] that collective legal interests are subject to an enforcement deficit, as the example of environmental law[43] explicitly shows. In the field of data protection, technical developments have contributed to undermining the assumption that privacy is fundamentally personal. Consequently, this assumption is no longer being reflected in the practice of data processing. As a result, data protection law has inevitably reached its limits.[44] It should be redesigned to broaden its perspective to include alternative theoretical foundations instead of further overburdening the current subjective approach. Such a redesign should also include questioning whether the collective legal interests described above should be guaranteed by data protection law or whether there are alternative instruments. The legal construction of the right to informational self-determination as an 'accessorial right' does not stand in the way of these considerations, but opens the door to collective elements. This is because this accessorial aspect could also refer to new legal rights that have yet to be developed, such as predictive privacy.

The protection of collective legal interests is well-known in the legal system and is, for example, much discussed in criminal and environmental law.[45] We could therefore conclude that not every protection of legal interests has to be directly reflected at the level of fundamental rights.[46] Nor does an individual understanding of fundamental rights prevent the further development of informational self-determination toward collective elements in other respects: some guarantees of freedom, such as freedom of assembly, can only be realised collectively.[47] These assupmtions are tranferable to data protection law.

3.5. Related concepts in ethical and legal debates

In order to legally define a protected good of predictive privacy, it is helpful to position the legal understanding in relation to the prevailing

[42] On collective rights see Michael Freeman, 'Are there Collective Human Rights?' (1995) 43 *Political Studies* 25; Leslie Green, 'Two Views of Collective Rights' [1991] 4 *Canadian Journal of Law & Jurisprudence* accessed 09 December 2022; Miodrag Jovanovic, 'Are There Universal Collective Rights?' [2013] 11 *Human Rights Review* 17.

[43] On collective rights to preserve the natural foundations of life: Shawkat Alam, 'The Collective Rights of Indigenous Peoples, Environmental Destruction, and Climate Change' in Erika J Techera et al Anastasia Telesetsky (eds), *Routledge handbook of international environmental law* (2nd ed, Routledge 2021).

[44] Daniel Solove, 'The Limitation of Privacy Rights' (2023) 98 *Notre Dame Law Review* 975.

[45] See for example: Jutta Brunnée, 'International Law and Collective Concerns: Reflections on the Responsibility to Protect', in Tafsir Malick Ndiaye and Rüdiger Wolfrum (eds), *Law of the Sea, Environmental Law and Settlement of Disputes* (Brill 2007) 35, 36.

[46] On collective dimensions of interferences within fundamental rights: Hannah Ruschemeier, 'Kollektive Grundrechtseinwirkungen. Eine verfassungsrechtliche Einordnung am Beispiel der Maßnahmen gegen die COVID-19-Pandemie' [2020] RW.

[47] On the collective dimension of Art. 19 Abs. 3 German Basic Law: Jan Oliva, 'Legal Persons from EU Member States and Their Entitlement to Fundamental Rights under the German Basic Law' [2011] 54 *German Y.B. Int'l L.*

understandings and discussions of individual-subjective, collective and 'group' rights, respectively. It is essential to first recognise privacy as a social value in order to derive further rights from it. Democratic societies and the rule of law that guarantee a human-centred protection of legal rights rely heavily on privacy.[48] In terms of individual rights, however, there is no fundamental right to a democratic society. An individual's right to a democratic society can only arise from the sum of individual rights, which, in interaction with other subjective and objective rights, transform the constitutional principles into legal reality.

3.5.1. Inferential privacy, group privacy and the 'right to reasonable inferences'

There are various proposals in the philosophical debate to conceptualise ethical problems related to predicted information. We use the concept of predictive privacy to distinguish between them.[49] *Loi* and *Christen* have used the term 'inferential privacy' to problematise potential privacy violations through predictions.[50] Unlike predictive privacy, however, *Loi* and *Christen* do not fully recognise the ethical problem of the 'prediction gap' (3.3) and use their term to capture a privacy violation only if the predicted information is based on logically valid inferences. The same objection applies to *Mittelstadt* and *Wachter's* concept of a Right to Reasonable Inferences:[51] predictive privacy aims to classify even the use of 'reasonably inferred' information as ethically and legally questionable and thus presents a *stronger* claim than the Right to Reasonable Inferences.[52] Similarly, predictive privacy differs from *Hildebrandt's* plea for a 'paradigm shift from data to knowledge protection' in the face of privacy violations through profiling.[53] In philosophy, knowledge is understood as true and justified belief. However, the violation of predictive privacy does not presuppose a prediction to be valid and thus qualify as knowledge. Therefore, the idea of 'knowledge protection' misses the point of predictive privacy.[54] Finally, predictive privacy also differs from the much-discussed concept of 'group privacy'.[55] This is

[48] Mireille Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20 *Theoretical Inquiries in Law* 83, 84.

[49] See 3.3 and Mühlhoff (n 6).

[50] Michele Loi and Markus Christen, 'Two Concepts of Group Privacy' [2020] 33 *Philosophy & Technology* 207.

[51] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI' (2019) *Columbia Business Law Review* 1.

[52] Mühlhoff (n 6).

[53] Mireille Hildebrandt, 'Who Is Profiling Who? Invisible Visibility' in Serge Gutwirth et al Sjaak Nouwt (eds), *Reinventing Data Protection?* (Springer 2009).

[54] Thus argues Mühlhoff (n 28) 4.

[55] Cf. Luciano Floridi, 'Open Data, Data Protection, and Group Privacy' (2014) 27 *Philos. Technol*. 1; Mittelstadt (n 32) 475; Linnet Taylor, Luciano Floridi and Bart van der Sloot (eds), *Group privacy: New challenges of data technologies* (Philosophical studies series, Springer 2017); Paula Helm, 'Group Privacy in Times of Big Data. A Literature Review' (2016) 2 *Digital Culture & Society* 137.

because the concept of predictive privacy does not tie ethical and legal concerns to the prerequisite that the invasion of privacy must occur through group-level predictions that affect a certain cohort of individuals identically and synchronously. Predictive modelling poses a new threat to privacy because it makes available a new realm of information — information that was never recorded, but is only predicted about the individuals concerned, effectively betting on the most likely outcome. The danger of abuse for this kind of information is independent of whether the algorithms proceed by virtual grouping of individuals or in some other way.

3.5.2. Legal consequences of group privacy

In terms of group privacy, it has become clear that the collective dimension of privacy consists of several layers: substantive and procedural rights; privacy as a group right, as a right to not be part of a group, as a right of the individual members of the group, or constructed as rights based on group membership both inside and outside the group. So far there is no coherent or consistent framework for group privacy, partly due to a lack of philosophical justification, which may lead to a weakening of fundamental and human rights.[56] The distinction between individual rights and collective or group rights is particularly difficult with regard to privacy protection when it is seen as a social and collective value for society.

According to the group privacy approach, predictive analytics make predictions about individuals based on their group membership and vice versa: assumptions about the group are fed by data from the associated individuals. Group privacy can thus be perceived as an individual right that individuals derive on the basis of their group membership or as a right of the group *per se*.

The threat to data protection and privacy posed by predictive analytics is therefore only partially addressed by concepts of group privacy. Moreover, the transfer of individual rights to group rights does not prevent people from becoming part of a group in the first place; it is also unclear which characteristics groups must have in order to be able to be holders of rights.[57] The grouping of predictive analytics through exploitation of collective data bases is neither based on information nor on consensus among stakeholders, as members know nothing about each other, and is also highly volatile.

With respect to predictive analytics, group privacy misses the real problem, which is that the privacy of any individual (even a single individual) can be violated by the data of many other individuals. In principle, every

[56] Johannes Morsink, ‘World War Two and the Universal Declaration’ [1993] 15 HUM 357, 397 (‘It seems, therefore, that the war, which prompted the writing of a Declaration with a set of universal and absolute values, did not provide a philosophy with which to defend that set’).

[57] Mittelstadt (n 32).

individual is affected, but the cause of the possibility of predictive privacy violation lies in the collective behaviour of many. In particular, the data basis for predictions about a particular individual is not limited to the data of that individual's own 'virtual group', but includes the data of all others, including individuals who are algorithmically distinguished from the particular case. Thus, the collective aspect of the privacy problem at hand is not that the algorithms may proceed with the formation of virtual groups (which does not apply to all predictive algorithms anyway), but that the collective behaviour of users, combined with insufficient regulation of the relevant technology, enables a privacy violation of potentially *any* individual (and group).

4. Predictive analytics and the GDPR

4.1. Classification of attacks within the context of data protection

A violation of predictive privacy occurs when personal information about an individual or group is *predicted*. Privacy violations through predictions constitute a comparatively less debated attack vector in privacy that is qualitatively different from other, more familiar types of attack, including intrusion and re-identification.[58] Unlike intrusive forms of privacy violation (hacking, breaking through encryption and security barriers, stealing data), where the target information is forcefully obtained, in the case of predictive privacy violations, obtaining the target information is not a process that breaks through manifest barriers. This is a feature that is shared between predictive violations of privacy and re-identification attacks. In comparison to re-identification attacks, however, predictive violations of privacy differ in two important ethical and legal aspects:

1. Predictive attacks potentially affect data subjects who are not themselves included in the training data on which the predictive model is based;
2. predictive attacks allow information to be estimated which the affected data subjects themselves never disclosed (because 'more sensitive' data are here derived from seemingly less 'sensitive' data).

We call this the dual *escalation structure* of predictive attacks. In comparison, re-identification attacks produce (1) only data about individuals included in the anonymised published dataset and (2) only data fields explicitly collected about those individuals. Predictive models, on the other hand, can (1) be applied to any third party individuals as soon as auxiliary data (input for the predictive model, e.g., usage data on a social media platform)

[58] See also Mühlhoff and Ruschemeier (n 11).

is known about them, and (2) estimate information that the individuals concerned have never provided to any third party or that they themselves may not know (e.g., disease predictions).[59] Another specific feature of predictive attacks is that they are typically scatter attacks that are potentially and simultaneously applied to many individuals (e.g., all users of a social media platform) by automated routines; this method is therefore not limited to targeted individual attacks on specific data subjects. For instance, as soon as a sufficient number of users of a large social media platform have explicitly stated in their profile that they smoke, the platform is able to train a predictive model that allows nicotine consumption to be estimated on the basis of usage data on the platform. This predictive model can then be automatically applied to *all* platform users, so that the platform is able to offer this information, which the majority of users have not actually provided, as a possible targeting criterion for advertisers.

4.2. Two processing steps: legal implications

It is generally plausible to assume that most predictive models are trained on anonymised mass data in order to circumvent application of the GDPR. As the scope of the GDPR requires the processing of personal data, it does not apply to the first processing step (training of a predictive mode) if the information used in this step does not have a personal reference. Nor does the GDPR need to be applied to the trained model itself as the product of the first processing step, because statistical or aggregated data do not have a personal reference to an individual (even if they only refer to a group of persons).[60]

While in theory, the GDPR may not apply provided that the data is anonymised, it is well known that in practice, even anonymised data sets bear the risk that individuals can be reidentified.[61] Given this blurred state of ‘anonymity’ of an individual in a data set, it is unclear how to exactly define the normative threshold for personal data in the context of predictive analytics.[62] If one assumes a low barrier, the majority of predictive models are likely to fall under the GDPR. This is especially true if one uses the persuasive three-element model put forward by the Article-29-Working Party.[63] According to

[59] Mühlhoff (n 6).

[60] Michèle Finck and Frank Pallas, ‘They who must not be identified—distinguishing personal from non-personal data under the GDPR’ [2020] 10 *International Data Privacy Law* 11.

[61] Paul Ohm, ‘Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization’ [2009] 57 *UCLA Law Rev.* 1701.

[62] Purtova argues that the distinction between personal and non-personal data is not suitable in the context of Big Data: 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. For privacy risks of LLMs: Hannah Brown et al. FAccT ’22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2280–2292; doi.org/10.1145/3531146.3534642 .

[63] Article 29 Working Party, Opinion 04/2007 on the Concept of Personal Data (WP 136) 01248/07/EN, 11.

this argument, personal data is identified as being one of three elements: content, purpose, or result. Irrespective of the purpose or result, a reference to a person can thus also be drawn from data that is derived from third parties, but refers to an individualised person.[64]

While the applicability of the GDPR to the first processing step may be disputed, when it comes to the second step (inference), the case is clear: If a predictive model is used to derive specific predictions about an individual case, this amounts to the processing of personal data and thus falls within the scope of the GDPR.

However, we argue that it is insufficient if the data protection mechanisms of the GDPR apply only with regards to the second step as this does not reflect how the first step (model creation) is already relevant to privacy — although this first step makes use of *collective* data. The first step is not captured by the concept of personal data and personal reference because model creation relies on the data of arbitrary third parties, not necessarily the data subject itself. There is therefore no 'data subject' in the sense of the GDPR referenced in model creation (training). The application of the model (inference) does affect an individual data subject if the inference is personal data, but the scope of affected persons in general is broader than the definition of 'data subjects', as the impact through training and validation goes beyond the individual application of the model.

4.3. Verifiability is not a condition of personal data

The verifiability of information has no influence on whether it qualifies as personal data. The argument that unverifiable information cannot be qualified as personal data assumes that the data subject's right of rectification provided by Art. 16 GDPR would not apply to unverifiable inferences,[65] and thus that the entire scope of application of the GDPR would similarly not be granted. This is a systematic fallacy; the rights of data subjects presuppose the application GDPR and not the reverse.

Regarding inferred data, it is argued that inferences are to be qualified as probability statements and thus cannot be verified, *ergo* they cannot be qualified as personal data. Indeed, this is true if only the material content of the prediction is considered, e.g., that person A will buy a house or commit a crime in the near future. Given the way the predictive model works technically, however, the purpose of the data processing cannot be the accuracy of the statement (output of the model) itself, but rather the efficient prediction. If this is taken as a basis, the right to rectification

[64] Finck and Pallas (n 60) 12.

[65] Gianclaudio Malgieri, 'Property and (Intellectual) Ownership of Consumers' Information: A New Taxonomy for Personal Data' (2016) PinG 133, 138 only wants to grant 'consumer rights' for inferred data and not data subject rights; this is also the argumentation of the defendant in BVwG Austria Partial Acknowledgement v. 26.11.2020 – W258 2217446-1/35E, BeckRS 2020, 51953 marg. no. 61.

cannot refer to the prediction result in the case of inferred data, but only to the statistical methodology. This clearly shows that there is a gap in legal protection with regard to the results of predictive models, which the Right to Reasonable Inferences (RRI) proposed by *Wachter* and *Mittelstadt* is intended to close.[66]

On the other hand, the specifically collective effects of predictive models illustrate the difference from the RRI[67] which addresses the matter of protecting the representation of the individual through inferred data (Right on How to Be Seen). The collective element of predictive data analytics is not grasped by new individually oriented legal rights such as the RRI. Nevertheless, the concepts of predictive privacy and collective approaches to data protection pursue the same goals as the RRI, which is to open a dialogue with data subjects and society about which processing practices are normatively acceptable.[68] However, individual rights necessarily encounter their limits when it comes to supra-individual effects; this thus requires a fundamentally different orientation, not least to compensate for structural asymmetries of power. The idea of collective elements of data protection may be mutually complementary with the concept of an RRI.

Unverifiable information can also have a significant impact on data subjects, as they cannot control how corresponding inferences are interpreted by data processors or third parties. The ECJ is interpreted as not seeing it as the purpose of data protection to guarantee correct decisions.[69] This is persuasive with regard to administrative procedures in the public sphere, where other mechanisms exist that ensure the correctness of the decision (legal protection, legal obligation of the administration, procedural rules based on the rule of law). The protective intent can only point in the direction that data protection law does not apply to inferred data if there are other contextual safeguards with regard to the rights of the data subjects. However, it does not necessarily follow that data protection cannot provide a remedy against misrepresentation if other safeguards are not available and the matter solely addresses the analysis of data for the purpose of producing the predictive model.

[66] Wachter and Mittelstadt (n 51).

[67] Ibid.

[68] Ibid, 1,92.

[69] This raises the question of which data subject rights are relevant de lege lata in predictive analytics and how effectively they protect against collective and individual interference. As already explained, the first step of building the predictive model (see 2.3 and 4.2) is not covered by the GDPR if it does not involve personal data, if the data is anonymised or cannot be attributed to the individuals whose information is processed: Wachter and Mittelstadt (n 51) 58.

5. Anonymisation and rights of the data subject in the context of predictive analytics

These conclusions raise the question of which rights of the data subject apply de lege lata against predictive analytics and how effectively they protect against collective and individual interference. As already explained, the first step of building a predictive model (see 2.3) is not covered by the GDPR where processing involves non-personal data, anonymised data, or data that cannot be attributed to specific individuals.[70] The GDPR also does not provide any measures against the use of personal data to make predictions about *third parties*; the understanding of data subjects' rights is always limited to their own data. Prerequisite for data subjects' rights is the existence of personal data, their allocation to an individual person, and the respective factual legal conditions of the subjective right.

5.1. Does anonymisation require a legal basis under the GDPR?

Considering the anonymisation of data, which might be involved in the creation of a predictive model, as a distinguished processing step which would then be subject to justification under Art. 6 GDPR does not solve the problem of collective data exploitation. First, it is unclear to which extent absolute and relative anonymisation techniques can provide sufficient protection against de-identification in big data analyses.[71]

Rather, how well anonymisation in the context of predictive analysis can protect against the privacy risk posed by privacy attacks become irrelevant, because prediction is not based on re-identification and may even target individuals who are not even included in the anonymised data set.

Second, and even worse: the communication strategies of large platform companies promise anonymisation and are implemented as a means of obtaining consent for the processing and analysis of users' personal data 'for statistical purposes', because users see their data protection and privacy needs as satisfied by the protection of their own anonymity. In the context of predictive privacy, this communication strategy is counterproductive, because the training of predictive models may occur on the basis of anonymised data; it is the promise of anonymisation that *enables* the production of prediction models.

[70] On the distinction between personal and non-personal data: Finck and Pallas (n 60). The right to erasure of the GDPR should not be covered by anonymisation: Alexander Roßnagel, 'Datenlöschung und Anonymisierung. Verhältnis der beiden Datenschutzinstrumente nach der DSGVO.' (2021) 11 ZD 188, 191; Bert-Jaap Koops, 'Forgetting Footprints, Shunning Shadows: A Critical Analysis of the 'Right to Be Forgotten' in Big Data Practice' [2011] *SCRIPT-ed.* 14.

[71] Ohm (n 61); Luke Munn, 'Staying at the Edge of Privacy: Edge Computing and Impersonal Extraction' [2020] 8 *Media and Communication* 270, 275; Daniel Solove and Paul Schwartz, 'The PII Problem: Privacy and a New Concept of Personally Identifiable Information' [2011] 86 *New York University Law Review* 1814, 1877.

From a legal perspective, the GDPR itself does not define anonymisation — which is not surprising at first, since anonymous data prima facie do not fall within the scope of the regulation. It is disputed whether the anonymisation of data is a form of data processing that requires legal justification and is thus subject to consent as required by Art. 6 (1) of the GDPR.[72] The right to data protection should not be affected by anonymisation, if there is no longer any personal data to protect. Even if one assumes that anonymisation requires consent under the GDPR, this reasoning focuses on the fact that individuals have an interest in maintaining their personal data within the current and original context of data processing.[73] Predictive privacy however, focuses on preventing illegitimate *secondary* data use. Predictive analytics in particular shows that processes can use anonymised data in various ways. The problem is not the process of anonymisation itself, but the subsequent use of the anonymised data for modelling. This secondary information processing is not covered by the GDPR in any way.

5.2. Rights of the data subject

The rights of the data subject under the GDPR are not able to solve the described problems and risks of predictive analytics. Even if the right of access provided by Art. 15 GDPR should apply and subsequently enable the deletion of one's own data, the follow-up question arises as to how many data subjects would have to exercise their right to deletion of their data in order to disable the predictive model. The same applies to the transparency obligations in Art. 13 and 14 GDPR; the information that one's own data is used to make predictions about others is not covered by the protective purpose of the norm, and is understood to be strictly individual.[74] Consequently, the right to rectification only relates to the methodology of the prediction.

According to the predominant understanding, Art. 22 (1) GDPR prohibits fully automated decisions.[75] This narrow understanding of the wording

[72] Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques (WP 216) 0829/14/EN, p. 5 subsumes anonymisation under 'further processing' of personal data. Schreurs/Hildebrandt/Kindt/Vanfleteren in Hildebrandt and Gutwirth (eds) (n 3) 249, 252 argue that rendering data anonymous does fit the definition of data processing under the GDPR and argue for a right to be informed about the anonymisation and a right to object that must be unconditional in the case of processing for indirect marketing purposes. UK Court of Appeal, *R v Department of Health*, para. 799 did not acknowledge anonymisation as a part of processing under the UK Data Protection Act 1998.

Against the requirement of a legal basis for anonymisation with considerable arguments: Gregor Thüsing and Sebastian Rombey, 'Anonymisierung an sich ist keine rechtfertigungsbedürftige Datenverarbeitung. (2021) ZD 548, 549.

[73] Schreurs/Hildebrandt/Kindt/Vanfleteren in Hildebrandt and Gutwirth (eds) (n 3) 249, 252.

[74] Instead of all: Gabriela Zanfir-Fortuna, 'Art. 13 Information to be provided where personal data are collected from the data subject' in Christopher Kuner et al. Laura Drechsler (eds), *The EU General Data Protection Regulation (GDPR)* (Oxford University Press 2020), 413, 415.

[75] Giovanni De Gregorio and Sofia Ranchordás, 'Breaking down information silos with big data: a legal analysis of data sharing', *Legal Challenges of Big Data* (Edward Elgar Publishing 2020) 226.

‘decision based solely on automated processing’ implies that this provision, originally intended as an ‘AI provision’, holds minimal practical relevance. Installing a *pro forma* human in the loop who merely confirms the output of the system makes it easy to circumvent the provision since the decision is no longer fully automated, instead the system is deemed to act in a decision-support capacity. Nevertheless, the measures required by Art. 22 (3) GDPR are one of the few provisions that refer to the result and not the procedure of data processing.[76] As procedural rights of data subjects, the provision at least guarantees the right to request the intervention of a person, the right to present one’s own point of view, and the right to contest the decision. However, these rights again only refer to fully automated decisions regarding the respective data subject and not to the collective data evaluation prior to the production of a predictive model. At this level, similar impairments within the meaning of Art. 22 (1) GDPR are difficult to determine; under the prevailing understanding, they are only likely to have legal effect in very few cases.

5.3. Distinction between metadata and inferences, Art. 9 GDPR

In terms of legal doctrine, the crucial blind spot of the GDPR stems not only from its limited application to individual, personal data, but to its neglect of the method of creating this data, which in the case of predictive models can logically only occur through interaction and comparison with the data of other persons.[77] Such a distinction between the acquisition and processing of information is only made to a limited extent in the systematics of the GDPR, for example, when Art. 9 GDPR differentiates between input data and metadata as well as the result of the data processing and thus pursues a risk-based regulatory approach.

Art. 9 (1) of the GDPR distinguishes two different prohibitions: Paragraph 1, first sentence, prohibits the processing of personal data from which certain sensitive categories of data, such as ethnic origin, can be inferred. The prohibition thus concerns the input data, which in turn do not have to be identical to the data of the processing result.

The process of inferring implies that sensitive data does not necessarily exist as part of the source data but potentially emerges in the course of data processing. This approach is taken to mitigate the risk associated with the possibility of inferring conclusions from general data that may also be considered sensitive data within the meaning of Art. 9 (1), first sentence of the GDPR. The specific differentiation in Art. 9 (1) is the

[76] Wachter and Mittelstadt (n 51) 1, 79.

[77] Ronald Leenes, ‘Reply: Addressing the Obscurity of Data Clouds’ in Hildebrandt and Gutwirth (eds) (n 3) 296, 297 on ‘correlatable humans’ not in context with the GDPR.

only reference in the GDPR to derived data. However, the distinction between source data and inferred data is not reflected in the broader protection regime and in particular, not in the rights of data subjects.

The second sentence of Art. 9 (1) GDPR prohibits the processing of certain sensitive data *per se*, e.g., biometric data, in terms of content; this means that only *potentially* sensitive data (like the categories listed in Art. 9 (1) (1)), which has been inferred from general data, is *not* addressed. The exact interpretation of the provision is disputed,[78] but the protective purpose is not determined by whether the content of the classification is correct with regard to the data subject.[79]

Simultaneously, the regulatory concept of Art. 9 (1) GDPR reveals considerable difficulties with the distinction between the different categories of data which underscores that the GDPR *de lege lata* is not able to address the complexities associated with big data.[80] One reason is that it is now potentially possible to infer sensitive information from almost any data, especially if one includes the boundless category of political opinions, c.f. Art. 9 (1) GDPR. Additionally, the fact that the data processor must be able to differentiate between general and sensitive data in the first place means the categorical distinction between sensitive and non-sensitive data in the GDPR is thus invalid, the special protection becomes obsolete.

It is disputed which criteria may be considered for the distinction between general and sensitive data. The intention of the data processing is discussed as a potential criterion for differentiating between sensitive and non-sensitive data, i.e., in the case of context-related information, there should be an intention to evaluate, which in turn can produce sensitive data.[81] Yet, if the intention is not clearly indicated by the objective circumstances of the data processing, the subjective element of the processor will not be verifiable in

[78] Ludmila Georgieva and Christopher Kuner, 'Art. 9 Processing of special categories of personal data' in Christopher Kuner et al. (eds), *The EU General Data Protection Regulation (GDPR)* (Oxford University Press 2020), 370, 371.

[79] ECJ Case C-465/00 *Österreichischer Rundfunk and Others* [2003] (ECLI:EU:C:2003:294), Rechnungshof, para. 75, stating 'To establish the existence of such an interference, it does not matter whether the information communicated is of a sensitive character or whether the persons concerned have been inconvenienced in any way.'.

[80] On the question whether inferences can be sensitive personal data: Wachter and Mittelstadt (n 51) 1, 70.

[81] Michael Matejek and Steffen Mäusezahl, 'Gewöhnliche vs. sensible personenbezogene Daten. Abgrenzung und Verarbeitungsrahmen von Daten gem. Art. 9 DS-GVO' (2019) ZD 551; *Schulz*, in Gola/Eichler et al. (eds.), DS-GVO, 2nd ed. 2018, Art. 9 DS-GVO, marg. no. 13; Thomas Petri, 'Art. 9 DSGVO' in Spiros Simitis, Gerrit Hornung Indra Spiecker gen. Döhmann (eds), *Datenschutzrecht: DSGVO mit BDSG* (C. H. Beck 2019), marg. no. 11; Wachter and Mittelstadt (n 51)1, 74; Thilo Weichert, '"Sensitive Date" revisited' [2017] 41 DuD; Georgieva and Kuner (n 78), 373 f.; Question No. 2 in ECJ Case C-252/21 *Meta Platforms and Others* [2021] (ECLI:EU:C:2022:704), dismissed by the Advocate general in his Opinion, Par. 41.

practice; there are also no normative indications for such an interpretation in Art. 9 (1) GDPR.[82]

Some argue that in case of doubt, the full scope of Art. 9 GDPR should not be applied.[83] This is not persuasive, as there are no normative indications for a fundamentally narrow interpretation. According to the internal systematics of the provision, the broad exemption rules of Art. 9 (2) GDPR instead support an equally broad understanding of the scope of application of Art. 9 (1).[84] Others argue from a supposed objective perspective, according to which there must be a significant probability or sufficient certainty in order to be able to derive sensitive data from the source data.[85] According to another view, only the specific processing context and the processing purpose should be taken into account, i.e.,he specific processing executed.[86] This line of argumentation shifts the scope of application of Art. 9 GDPR in terms of its temporal scope: if it is important that source data are actually used for inferring sensitive data, it may only be determined during the processing of the data itself. If one follows this line of reasoning, the mere possibility that sensitive data *may* be inferred from general data becomes irrelevant. It remains unclear whether this specific reasoning could still be applicable to data processing via predictive analytics. This is because, in such instances, the data processing interlinks a range of information, rendering the majority of the personal data processed theoretically usable as source data for inferring sensitive data. The case of predictive analytics argues for a focus on the specific processing *context* to determine whether the appropriateness and probability of use is sufficient. Yet, in the context of big data, these requirements will probably always be considered fulfilled in light of the technical mode of operation by which the processing occurs.

The regulatory approach of Art. 9 GDPR, which distinguishes between general data, e.g., the creation of the model, and inference, e.g., the results of the model when it is applied to a single case, can be transferred to predictive models in terms of the protective purpose of the norm: the metadata used as part of the collective data evaluation during the creation of the predictive model form the basis for individual-related inferences. The non-sensitive source or proxy data, from which the sensitive data within the meaning

[82] Petri (n 81), marg. no. 12.

[83] Sebastian Schulz, 'Art. 9 DS-GVO' in Peter Gola et al. (eds), *Datenschutz-Grundverordnung: VO (EU) 2016/679: Kommentar* (2nd ed. C.H.Beck 2018), Art. 9 DS-GVO, marg. no. 13.

[84] 'Neither narrow nor broad interpretation' Thilo Weichert, 'Art. 9 DS-GVO' in Jürgen Kühling Benedikt Buchner (eds), *Datenschutz-Grundverordnung/BDSG: Kommentar* (3rd ed. C.H. Beck 2020), mag. no. 22.

[85] Advocate General ECJ (Fn. 62) Par. 38; Marion Albers and Raoul-Darius Veit, 'Art. 9 DS-GVO' in Stefan Brink Heinrich A Wolff (eds), *Beck'scher Online-Kommentar Datenschutzrecht* (42. ed. C. H. Beck) marg. no. 21 f.; Petri (n 81) marg. no. 12; Alexander Schiff, 'Art. 9 DS-GVO' in Eugen Ehmann Martin Selmayr (eds), *Datenschutz-Grundverordnung: Kommentar* (2nd ed C. H. Beck 2018) marg. no. 13; Christian Bergauer, 'Personenbezogene Daten, Begriff und Kategorien' in Rainer Knyrim (ed), *Datenschutz-Grundverordnung* (2016) 43, 60.

[86] Matejek and Mäusezahl (n 81) 551, 553.

of Art. 9 (1) GDPR can emerge, correspond to collective data analysis when transferred to predictive models. However, both can result in derivations that concern individuals or fall into the category of sensitive data.

This collective dimension means consent can never be a viable basis for corresponding predictive decisions, since the individual person can only consent for themselves and not for all other data subjects. A prohibition of the further processing of the data should arise from predictive analytics and the model output, even if data subjects have consented to it. This is because the predictive model output can only arise from collective data analysis. As the case of Meta illustrates, this model output is generated based on predictions about users who have not explicitly consented to the processing of their data, or even from the data of non-users. Therefore, in practice, the category of 'data subjects' in the sense of data protection goes far beyond the scope set out in the GDPR.[87] However, the way predictive analytics works means that there can be no real differentiation between source data and potentially particularly sensitive data if everything can in principle be derived from large data sets.

The difficulty inherent in differentiating between data covered by Art. 9 (1) GDPR and that which is not has reached the ECJ, which addressed this prominent issue in August 2022.[88] Following a request for a preliminary ruling by the Administrative Court of Vilnius, the ECJ qualified name-specific data relating to the partner of an applicant for a position within the public administration which had to be revealed during the application process as sensitive data under Art. 9 (1) GDPR. The Court decided that the name of a spouse, partner or cohabitee has the potential to reveal the sexual orientation of the applicant.[89] Furthermore, the court set only low requirements for 'revealing' sensitive data: accordingly, an 'intellectual operation involving comparison or deduction'[90] is a sufficient condition for extending the special regime of protection envisaged for the protection of sensitive data to personal data, which are not inherently sensitive. However, as this judgement did not relate directly to predictive analytics or big data, the distinction remains unclear. The ECJ dismissed the purpose-based limitation which relies on the intention of the data processor in favour of a contextual approach, without explicitly specifying the criteria for identifying potentially sensitive personal data. The decision also does not clarify whether this interpretation can be applied to other contexts, especially outside the state sphere, e.g., to big data analyses by private companies.

[87] Emőke-Ágnes Horvát et al, 'One plus one makes three (for social networks)' (2012) 7 PloS one e34740; which also clearly contradicts the principle of confidentiality and transparency, Art. 5 (1) a, f GDPR.
[88] ECJ C-184/20, Vyriausioji tarnybinės etikos komisija [2022] (ECLI:EU:C:2022:601).
[89] Ibid, par. 119.
[90] Ibid, par. 120.

6. Jurisprudence on inferred data and predictive analytics

6.1. ECJ and inferred data

Predictive models produce inferred data, they generate predictions about individuals (data subjects), cases, or events. Regardless of whether the process of creating these models is covered by the GDPR, there is a discussion about whether and to what extent the protective effect applies to inferred data.[91] This would require that inferred data are to be classified as personal data.

The case law of the ECJ is somewhat reticent on these questions, although reference is made elsewhere that the ECJ has commented on the qualification of inferred data.[92] As part of a preliminary ruling, the court was asked to decide on whether abstract legal assessments made in connection with a person, i.e., the 'conclusions' of the data processing itself, qualified as personal data.[93] The ECJ ruled that the legal analysis qualified as a process and thus not as personal data. On the other hand, the facts underlying these conclusions were qualified as personal data because they were individually related to the person. This differentiation, between the process of legal analysis and the underlying personal information of the decision, is sensible. That said, no conclusions can be drawn from this decision as to whether the court qualifies inferred data as personal data.

In its reasoning, the court referred only to the context of the inferred data, not its structural properties. In this case, the 'legal analysis' in question was part of the administrative procedure for granting a residence permit, involving the preparation of the facts and legal assessment by the competent authority for the purpose of preparing a final decision; a process specifically covered by the procedures of the right to information under Art. 15 GDPR.[94]

The ECJ's reasoning on the protective purpose of the GDPR in the specific situation of administrative procedures was clear: the right of access is intended to enable data subjects to demand the correction, deletion, or blocking of data. However, this right cannot relate to the legal analysis of the public authority,[95] because this analysis is based on the administration's (legally required) application of the law and not on the protection of personal

[91] Wachter and Mittelstadt (n 51), 1; 47 assume limited protection for derived data.

[92] Ibid, 1, 29.

[93] ECJ Case C–141/12 *YS and Others* [2014] (ECLI:EU:C:2014:2081) and ECJ Case C–372/12 *M and S* [2012] (ECLI:EU:C:2014:2081). Following: AG München, Partial judgment of 4.9.2019–155 C 1510/18.

[94] The initial impetus for the procedure was the change in the administrative practice of no longer providing the data subjects with the legal analysis upon simple request, but instead a summary of the personal data contained and processed and the bodies that dealt with it., see ECJ Case C–141/12 and C–372/12 (ECLI:EU:C:2014:2081).

[95] ECJ (n. 93), marg. no. 45.

data. Indirectly, this is also a question of the separation of powers; an authority's interpretation of the law may be reviewed within the framework of an appeal procedure but not by the exercise of an individual's right to information. Thus, the case cannot provide for the conclusion of a general statement on the qualification of derived data.

This also applies to the *Nowak* case[96] in this context. The case concerned an examinee's right to information regarding the script of a failed admission examination. The ECJ classified the examiners' comments as personal data, since they constituted information about the examinee themselves, which also includes opinions and assessments.[97] Although the court rejected the application of the right of rectification,[98] this does not imply reduced protection for inferred data, because the reasoning was again not system-based, but purpose-oriented and contextual. In the case of examination performance, it is logically contrary to the purpose of the data processing — and the examination itself — if answers can be corrected later. Therefore, this decision does not reduce the protective effect of the GDPR for inferred data which qualify as personal data. However, this only applies to the individual components of identifiable data subjects.

The rulings of the Austrian Federal Administrative Court and the Supreme Court that information on probability statements about certain affinities of a person can be considered personal data within the meaning of the GDPR, also follow this line of reasoning.[99] Before the GDPR came into force in Austria, the right to information under Section 32 of the GDPR 2000 could only be asserted by the data protection authority and not by the data subjects, but this exclusive competence can no longer apply against the background of Art. 15 of the GDPR.

The Administrative Court clearly ruled that the processing of personal data from which party affinities are derived as well as probability values constitutes personal data.[100] The methodology and the result are persuasive, as the court relied on the coherently developed characteristics put forward by the Article 29 Data Protection Working Party for defining personal data: content — purpose — effect.[101] The comparison with official statistics was also an interesting element of the decision: according to the court, no connection between an individual's socio-demographic data and their interest in certain parties was established; instead, the assessment of party affinity related to avoiding scatter losses in advertising, thus

[96] ECJ Case C-434/16 *Nowak* [2017] (ECLI:EU:C:2017:994).
[97] ECJ (n. 93), marg. no. 34.
[98] At that time still Art. 12 b RL 95/46 before the entry into force of the GDPR, ECJ (n. 93)., marg. no. 52.
[99] BVwG Austria Partial recognition v. 26 November 2020 – W258 2217446-1/35E, BeckRS 2020, 51953; OGH Wien, Urteil vom 18.2.2021–6 Ob 127/20z (OLG Linz), BeckRS 2021, 20609.
[100] BVwGÖ (n. 98), marg. no. 55 ff.
[101] BVwGÖ (n. 98), marg. no. 60 ff.

required a link to individuals to achieve its purpose.[102] The Administrative Court specifically did not conclude from the inapplicability of Art. 16 GDPR that probability statements are excluded from the entire scope of the GDPR. A view to the contrary is systematically unconvincing and, moreover, factually inaccurate: the purpose of data processing is not the correctness of the assignment, but solely its methodologically substantiated assessment.[103]

Probability predictions as inferred data of predictive models thus fall within the scope of the GDPR. However, this does not sufficiently consider the collective dimension of modelling, as data protection rights and the GDPR as a whole only refer to the data of the individual person and not to their effects on third parties.

6.2. The Meta-case

In its recent ruling, the ECJ identified several noteworthy conclusions, in particular confirming competition law as an effective enforcement tool for data protection law.[104] The decision confirms many of the arguments made in the literature against the risks of predictive analytics, but does not solve the fundamental problems.

The Court made a passing reference to the problems of distinguishing between the different categories of personal data raised here, acknowledging that it was no longer factually possible to do so.[105] The ECJ also rightly assumes that for the special categories of personal data it does not matter whether the information is correct or not.[106] The scope of Art. 9 (1) GDPR is interpreted broadly, which, in combination with the assumption that special categories of personal data also exist in a mixed data set ('infection effect'), leads to the conclusion that the processing of large data sets cannot in fact be designed in a legally compliant manner. This is because the ECJ relies on consent here and assumes that it is not already excluded due to Meta's dominant market position (para 140 et seq.) As a result, this is understandable, as the problems of consent in the digital context are not only fed by the dominant market position of processors, but above all by the sheer flood of information. These circumstances continue to make consent problematic.[107] The ECJ's comments on the processing basis of the legitimate interest, Art. 6 (1) (f) GDPR, are convincing: the presumption of legitimate interest requires a balancing of interests between data subjects

[102] BVwGÖ (n. 98), marg. no. 65.
[103] BVwGÖ (n. 98), marg. no. 71.
[104] ECJ C-252/21, see also: Hannah Ruschemeier, https://verfassungsblog.de/competition-law-as-a-powerful-tool-for-effective-enforcement-of-the-gdpr/.
[105] ECJ C-252/21, par. 89.
[106] Ibid, par. 69.
[107] Ruschemeier (n 12).

and processors, considering structural violations of the GDPR, the dispersion of processing in terms of data and data subjects, as well as the exploitation of collectively generated data bases.

With regard to the dangers of predictive analytics, the ECJ moves along the lines of the GDPR and unsurprisingly does not fundamentally question it. However, its comments clearly show the limits of data protection. The court understood the Metas business model precisely and stated this very clearly:

Furthermore, the processing at issue in the main proceedings is particularly extensive since it relates to potentially unlimited data and has a significant impact on the users, a large part — if not almost all — of whose online activities are monitored by Meta, which may give rise to the feeling that their private life is being continuously monitored (para. 118).

This decision explicitly acknowledges the problem of power asymmetry raised here, and the challenges of prediction and collectivity remain. Finally, the ruling explicitly recognises that data is power. Its impact will nevertheless remain limited for the time being, and the referring court will have to decide anew, considering the questions referred. So far, only the German competition authorities have dared to act against Meta to this extent.

7. Predictive analytics and the Digital Services Act (DSA) and Digital Markets Act (DMA)

7.1. DSA

The problems addressed in this paper are not sufficiently solved by the new European legislation on digitalisation.[108] From the outset, it must be noted that the DSA[109] is intended to regulate systemic risks and therefore would be well suited to address the kind of collective data exploitation and manifestations of data power described in this paper. The DSA does not, however, include very substantive provisions to protect privacy or other fundamental rights against the very specific threats arising in the context of aggregate data

[108] The Data Governance Act (Regulation on European Data Governance COM(2020) 767 final), the Directive on Copyright and Related Rights in the Digital Single Market (Directive 2019/790) and the drafts for the Data Act (proposal on harmonised rules on fair access to and use of data, procedure 2022/0047/COD), the Artificial Intelligence Act (Regulation laying down harmonised rules on Artificial Intelligence, procedure 2021/0106/COD), the Regulation on the European Health Data Space (procedure 2022/0140/COD) are not included in this analysis, as their impact on predictive analytics is of secondary importance or not yet foreseeable.

[109] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC.

usually collected by digital platforms. Data protection and privacy are mentioned (Art. 28 (1), 38 (1b), 46 (2)), but are not at the centre of the regulatory intention of the DSA.[110] Art. 2 (4) (g)) states that generally the DSA should be without prejudice to other acts of Union Law, including the GDPR. But the devil is in the details: although Art. 25 (1) DSA contains useful prohibitions against manipulation, deception, and preventing individuals from making free and informed decisions, Art. 25 (2) then excludes these prohibitions from application to practices within the scope[111] of the GDPR, and thus to all the consent-based data processing operations that facilitate data aggregation and predictive analytics as discussed in this paper.

The regulatory regime of the DSA is not sufficiently rooted in the theoretical foundations of the socio-technological ecosystems of the internet, nor in the individual and collective rights of the users most affected by the social inequality and new power hierarchies resulting from societal threats posed by platform business models. As a mainly procedural framework,[112] the DSA delegates too much of the power to define the substantive rules and criteria (When is content illegal, offensive, racist? What constitutes hate speech, fake news, etc? When are the limits of free speech violated?) to the self-regulation of the large internet platforms and search engines. Neither does the DSA provide legal tools to those affected.[113] Furthermore, the DSA is extremely focussed on user-generated content, rather than addressing the systemic questions behind the problematic developments of the internet, e.g., why Fake News and other malicious content can be spread digitally with such ease, and what role design, moderation, economic interests, and concentrations of power of platform operators play. Although the criteria of systemic risks (c.f. Art. 34 DSA) go beyond the individual affected, the broad obligations of the very large online platforms to conquer these systemic risks give them more power to self-regulate with seemingly only marginal opportunities for external intervention by public authorities, civic bodies, or affected communities.[114] The most significant shortcoming of the DSA is not recognising the platforms' own business model based on aggressive data extraction as a systemic risk.[115]

[110] Recitals 71 and 103 mention the protection of privacy with regards to the protection of minors and the voluntary codes of conduct.

[111] Art. 25 (2) DSA also excludes practices under the Unfair Commercial Practices Directive (2055/29/EC).

[112] Therefore, the DSA is not the 'new constitution' of the internet as discussed in <https://en.alexandrageese.eu/video/europe-calling-dsa-deal/> accessed 12 December 2022.

[113] Hannah Ruschemeier, 'Re-Subjecting State-Like Actors to the State' in Heiko Richter, Marlene Straub Erik Tuchtfeld (eds), *To Break Up or Regulate Big Tech? Avenues to Constrain Private Power in the DSA/DMA Package* (2021) (Max Planck Institute for Innovation and Competition, Research Paper No 21-25) 49.

[114] Ibid, 51.

[115] On this in detail: Hannah Ruschemeier, Art. 34 DSA in Spindler et al. (eds.) *Recht der elektronischen Medien* (5th ed) (CH. Beck 2024, forthcoming).

Many digital platforms use business models that leverage predictive analytics, relying on the invalid legal basis of user consent[116] to facilitate data collection and aggregation. These malicious business practices are barely addressed by the DSA. While the DSA prohibits the manipulation and deception of users (Art. 25 (1) DSA[117]), this provision is insufficient in at least three ways in face of the substantive threats debated in this paper. First, while the DSA does address platforms as actors in the relevant norms, it seems unclear in what way the requirements, e.g., of Art. 25 DSA, also apply to the content and services embedded in the platform's website, like, e.g., Google Ads. Secondly, it seems unclear in what way these provisions also hold in relation to non-users (that is, persons who are not registered on the platform but still have their data processed, e.g., as members of contact lists, entries in phone books etc.). Thirdly, the provisions related to advertising and recommender systems, Art. 26, 27 DSA are procedural norms of transparency and do not establish any rights of users or prohibitions concerning secondary use of aggregate data.

At first sight, individual provisions of the DSA might even seem to be stricter than the GDPR: Art. 26 (3) prohibits 'advertisements to recipients of the service based on profiling'[118] using 'special categories of personal data'.[119] As a result, the exceptions of Art. 9 (2) GDPR do not apply, that is, the consent of the data subject cannot override the Art. 26 (3) prohibition. However, the prohibitions in Art. 26 (3) are limited in their scope since they apply only to advertisements shown to the users of digital services that qualify as online platforms under the DSA. The same online platforms are not hindered by the DSA from providing targeted advertising based on profiling to other websites, apps, and services that do not qualify as a 'platform' under the DSA.[120] This means in particular that the DSA does not prevent the training and production of predictive or profiling models from

[116] 97% of the 75 most popular websites use 'dark patterns', manipulating users into consenting to the processing of their personal data: https://open-evidence.com/2022/06/10/behavioural-study-on-unfair-commercial-practices-in-the-digital-environment-dark-patterns-and-manipulative-personalization/ (accessed 12 December 2022); Karen Yeung, ''Hypernudge': Big Data as a mode of regulation by design' (2017) 20 *Information, Communication & Society* 118 describes these 'hypernudges' as the biggest threat for fundamental rights by big data; on the question why consent is often invalid in digital environments: Hannah Ruschemeier, 'Privacy als Paradox?' in Michael Friedewald Alexander Roßnagel (eds), *Künstliche Intelligenz, Demokratie und Privatheit* (Nomos 2022).

[117] Which prohibits actions that 'deceive or manipulate the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions'.

[118] As defined by the GDPR. According to Art. 4 GDPR 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

[119] Referred to in Art. 9 (1) GDPR.

[120] For very large online platforms, Art. 39 DSA requires additional online advertising transparency rules, but no substantive provisions.

the data of platforms, but only prohibits their utilisation in certain contexts. Neither does the DSA cover the privacy risks associated with large language models, as those models and their applications do typically not meet the criteria of a platform.[121]

When it comes to the exploitation of collective data, the distinction in Art. 9 (1) GDPR between general data and sensitive data is obsolete from our point of view. In recent case law, the CJEU has interpreted Art. 9 (1) very broadly,[122] rendering the difference between general data and sensitive data unmanageable in practice when it comes to big data. Nevertheless, the DSA adopts this distinction by directly referring to Art. 9 (1) GDPR in Art. 26 (3).

This problem is more effectively solved by the stricter provision of Art. 28 (2) DSA. This norm is a step in the right direction and explicit recognition of the problems of targeted advertising. According to this prohibition, platform providers shall not display any advertising on their interface that is based on profiling pursuant to Art. 4 (4) GDPR using personal data when they are aware that the recipient of the service is a minor. It is not clear whether the prohibition of Art. 28 (2) DSA also has a broader scope than Art. 26 (3) DSA since Art. 28 refers to 'online interfaces', whereas Art. 26 (3) DSA addresses the 'presentation to recipients of the service' which could be the platform itself or the online interface.

Overall, the DSA fails to address data power as an all-encompassing principle of the social and economic dominance held by digital service platforms. The DSA addresses this problem implicitly at best, by partially sanctioning some of data power's effects.

7.2. DMA

The DMA aims to protect competition in the digital economy in the Union and is thus a response to the structural characteristics of digital markets and the market power of some platforms operating in these markets, as well as to the limited impact of competition law in this area. Thus, the excessive economic power of a few gatekeepers, cf. Art. 3 DMA, leads to significant imbalances in bargaining power, which is why market processes in the digital economy often fail to ensure fair outcomes. It remains to be seen whether the ECJ will continue to develop its jurisprudence on the interaction between competition and data protection law. Enforcement of the DMA is the sole responsibility of the Commission, Art. 20 f. DMA, with initiatives by national competition authorities taking precedence. Whether the

[121] Hacker et al. FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency June 2023 1112 https://doi.org/10.1145/3593013.3594067 ; Hannah Ruschemeier https://verfassungsblog.de/squaring-the-circle/.

[122] ECJ (n. 87).

Commission will use the DMA to crack down on data power, even if it does not affect competitors, is an open question. In any event, the Commission has provisionally concluded that Meta has imposed unfair trading conditions on competing online classifieds services. Should this approach be pursued further, the DMA could prove to be a very effective tool to enforce data protection law.

As the business models of the very large online platforms are based on the systematic exploitation of collective databases, it may be difficult to reconcile the role of the European Commission with the policy decisions or legislative proposals made by the same institution in the area of data protection law or other related areas. For example, in the area of data transfers to third countries, in particular to the US, there is an overlap between different policy objectives of the Commission: on the one hand, the promotion of international trade and the strengthening of the Union as a business location and, on the other hand, the protection of the fundamental rights of Union citizens, as repeatedly demanded by the ECJ. Similar conflicts are conceivable in the area of supervision of the very large online platforms.

8. Conclusion

Predictive analytics is ubiquitous, but the collective impact of predictive analytics is rarely the subject of debate. This is partly because the impact is less obvious, as the workings of predictive models are opaque and technically complex. As we have shown, neither the individual regulatory framework of the GDPR nor the DSA are adequately equipped to address the risks of predictive analytics with respect to the collective dimensions of data exploitation. We employed the concept of prediction power to name the specific manifestation of data power in the context of predictive analytics. We discussed how prediction power leads to new regulatory questions insofar as the analytic models that specifically threaten privacy can only be deployed by a limited number of identifiable institutions, predominantly private companies and, to a certain extent, state actors. The need for protection of predictive privacy begins where prediction power starts to emerge (step of model creation), and not only in the manifest exercise of this power in individual cases (inference step). We proposed a conceptual, ethical, and legal framework on the issues of predictive analytics and privacy to enable researchers and policy makers to collectively impact the debate on the regulation of digital technologies.

We also argued that collective elements of privacy and data protection cannot, and should not, replace individual rights protection, which is more important than ever in the age of ubiquitous data analysis. Ideally, they should complement each other. Awareness and transparency of causal structures are needed, and mandatory publication of predictive models is a reasonable first step. Creating a collective perception of rights needs to be part of the strategy against the collective exploitation of data through predictive analytics characterised by informational power asymmetries. The regulatory framework for power asymmetries should acknowledge prediction power as a relevant market factor. Only enforcement will show whether the current legislation effectively addresses the risks of predictive analytics. In any case, a theoretical foundation for understanding is needed. Many of the challenges are also reflected in the applications of generative AI. This in turn requires theoretical elaboration to unite the perspectives of different disciplines.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Prof. Dr. Rainer Mühlhoff is Professor of Philosophy and Head of the Research Group on the Ethics and Critical Theories of Artificial Intelligence at the Institute for Cognitive Science at the University of Osnabrück. His research focuses on ethics, social philosophy, philosophy of technology and media studies of the digital society.

Prof. Dr. Hannah Ruschemeier is junior professor of public law, data protection law and the law of digitalisation (tenure-track W3) at the University of Hagen, board member of RAILS e.V. and co-editor of the journal Legal Tech. Her research combines traditional public law issues with the challenges of digital transformation.