

Consumer-friendly scoring



Recommended citation:

SVRV (2018) Consumer-friendly scoring. *Gutachten des Sachverständigenrats für Verbraucherfragen*.

Berlin: Sachverständigenrat für Verbraucherfragen.

Berlin, October 2018

Report of the Advisory Council for Consumer Affairs

ISSN 25100084

Published by:

Advisory Council for Consumer Affairs

at the Federal Ministry of Justice and Consumer Protection

Mohrenstraße 37

10117 Berlin

Telephone: +49 (0) 30 18 580 0

Fax: +49 (0) 30 18 580 9525

email: info@svrverbraucherfragen.de

Website: <http://www.svr-verbraucherfragen.de/en/>

Design: Atelier Hauer+Dörfler GmbH, Berlin

Printed by: Druck- und Verlagshaus Zarbock GmbH & Co. KG, Frankfurt am Main

© SVRV 2018

Consumer-friendly scoring



Sachverständigenrat
für Verbraucherfragen

Acknowledgements

Part of the mission of the Advisory Council for Consumer Affairs (SVRV) involves incorporating new research findings and practical experience into the drafting of its publications. In order to meet this requirement as comprehensively as possible, the SVRV has created various serial publications – reports, working papers and commissioned studies. It also stages specialised events and engages in public and non-public conversations with representatives of the academic and business communities and of civil society. With the aid of representative and non-representative surveys, public opinion informs the work of the SVRV, as do the legitimate interests of business enterprises. Without the assistance and cooperation of these individuals and institutions it would not have been possible to compile this report.

The SVRV thanks all of the staff of its office for their outstanding work on the preparation of this report. We extend a special word of thanks to the research staff – Johannes Gerberding, Dr Christian Gross, Dr Ariane Keitel and Sarah Sommer – as well as to Thomas Fischer, head of the SVRV Office, and to the temporary deputy head of the SVRV Office, Stefan Kubat.

Dr Felix Rebitschek, Jasmin Ghalib und Juri Ritz assisted the lead authors of the report, Gerd Gigerenzer and Gert G. Wagner, in their work. Our thanks go to them too.

This report is also based in part on material that has been published in the SVRV Working Papers series.

The SVRV thanks the authors of the working papers *Verbraucher-Scoring aus Sicht des Datenschutzrechts* ('Consumer scoring in the light of data protection law') and *Dokumentation einer empirischen Pilot-Studie zum Wissen über und zur Bewertung von Verbraucher-Scoring* ('Documentation of an empirical pilot study on awareness and assessment of consumer scoring').

We also thank the authors of the study *Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren* ('Technical and legal reflections on algorithmic decision-making processes') – Professor Georg Borges, Dr Matthias Grabmair, Daniel Krupka, Professor Burkhard Schäfer, Professor Erich Schweighofer, Professor Christoph Sorge and Bernhard Walzl of the Specialist Group on Legal Informatics of the German Informatics Society.

Before and during the preparation of this report, members of the SVRV and of its Bureau held numerous conversations with representatives of the academic and business communities and of NGOs. Our thanks go to the following (in alphabetical order):

Professor Andreas Beyer (University of Cologne), the Bundesanstalt für Finanzdienstleistungsaufsicht (Federal Financial Supervisory Authority), the Bundesversicherungsamt (Federal Insurance Office), Wolfie Christl (Cracked Labs – Institut für kritische digitale Kultur), Dacadoo AG, Generali Deutschland AG, Professor Justus Haucap (University of Düsseldorf), Lorena Jaume-Palasi (Algorithm Watch), Professor Genia Kostka (Free University of Berlin), infoscore Consumer Data GmbH, Professor Sergio Lucia (Technical University of Berlin), Professor Stefan Lessmann (Humboldt University, Berlin), Professor Mario Martini (German University of Administrative Sciences in Speyer), Professor Steffen Mau (Humboldt University, Berlin), Professor Klaus-Robert Müller (Technical University of Berlin), Walter Palmetshofer (Open Knowledge Foundation e.V.), Professor Eberhard Sandschneider (Free University of Berlin), Schufa Holding AG, Professor Wolfgang Schulz (Hans Bredow Institute and Alexander von Humboldt Institute for Internet and Society), Arne Semsrott (Open Knowledge Foundation e.V.) and Sparkassen DirektVersicherung AG.

A total of 75 business enterprises or health insurance funds took part in a market study covering the segments of credit, driver and health scoring. The SVRV offers its sincere thanks to them for their cooperation as well as to those who took part in a representative survey on public awareness and acceptance of scoring. The survey was conducted by infas, the Institute for Applied Social Sciences, with Ms Janina Belz as head of project. We are grateful to all of these important partners.

Dr Philipp Hacker of the Berlin Social Science Research Centre (WZB) and Christin Schäfer of the consultant firm acs plus: data with care have earned our gratitude not only for their critical appraisal of the draft report in the form of an independent peer review but also for many important suggestions that have been incorporated into the final version.

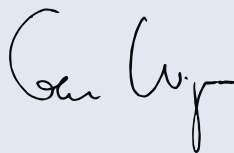
Please note that this text represents a translation of the original report published in German, excluding the annex of the original report. Therefore, any reference to the annex shown in this text refers to the annex of the original report. The language in the text of this report is intended, in principle, to be gender-neutral. For the sake of better readability, however, we have refrained from continuously referring to both sexes separately.



Lucia Reisch



Gerd Gigerenzer



Gert G. Wagner

Executive Summary – Recommended actions for consumer- friendly scoring

1. Making scoring comprehensible for consumers

1. The Advisory Council for Consumer Affairs recommends that data protection authorities operationalise the *comprehensibility requirements* set out in the GDPR (cf. Article 15 para. 1 letter h) for scoring and score-based business processes. Comprehensibility should be measured according to the standards relevant to the average consumer. Where scoring entails a level of complexity that is no longer comprehensible to the individual consumer, measures should be taken to ensure that scoring processes can be understood not only by supervisory authorities, but, at the very least, by consumer bodies and non-state actors as well.
2. Scoring services should release clear and comprehensible information for consumers about the main criteria used to score them and, preferably, how these variables are weighted. Trade secrets, of course, must remain inviolable. The definition of which variables are considered crucial for consumers cannot be left exclusively to lawmakers: this task should additionally fall within the remit of consumer organisations, or, alternatively, the “market watchdogs” of Germany’s consumer advice centres. At any rate, full disclosure to supervisory authorities of scoring systems and their attributes is a must (see page 5 of the Advisory Council’s *Digital Sovereignty* report).

Some members of the Advisory Council advocate further-reaching transparency. They believe that all scoring variables should be disclosed to the consumer and that the relative weighting of each component should be indicated in the calculation of the score. To this extent, any interests on the part of scoring services and users in maintaining secrecy would take second place to the consumer’s interest

in receiving information. At the same time, the trade secret of how a scoring system has been developed and programmed would be maintained.

3. However, disclosure alone will not necessarily give consumers a better understanding of how scoring works. This will require a variety of measures, which include: providing examples of consumer scores and how they are tiered according to different variables; the production of visual teaching aids (e.g. by consumer organisations); general efforts to raise scoring-related competence among consumers. Any assessments of how comprehensible scores are to consumers should be based not only on expert opinion but on *empirical evidence*.
4. Consumers already have a right to tailored and meaningful written information whenever they are scored (see Article 13 para. 2 letter f, 15 para. 1 letter h GDPR). However, this right has not yet been set out in *more concrete terms*. Companies, supervisory authorities and consumer organisations should work together to develop *standards for scoring services*, which would help guarantee relevance and comprehensibility. The Advisory Council further recommends informing consumers of how their personal score is to be interpreted against the distribution of score values among the population as a whole (e.g. does my score put me in the “upper third”?).
5. Prompt, free-of-charge notification should be provided – or at least offered as an option for consumers – in the event of *major changes to a person’s score* (e.g. if the person slips into a lower category). Naturally, there are certain limitations to this: in order to register a change in score, scoring services would have to retain historical score values. There are many practical applications (such as fraud recognition or determining possible payment modalities) for which this option will not be available. At banks and insurance companies, scores are calculated on an ad-hoc basis. This means that no score

history is maintained, and potential changes are not apparent at the time the next “event” is registered. This proposal can therefore be implemented only at institutions where data collection is ongoing, e.g. credit scoring services and the Federal Motor Transport Authority in Flensburg (with its “Register of Driver Fitness”, which already sends out such notifications).

2. Fostering knowledge and competence

As recommended in the Advisory Council’s Digital Sovereignty report, NGOs, consumer protection organisations and consumer protection projects should provide education on basic issues related to scoring in all its manifestations, as well as on the use of scoring in specific fields of business.

1. For this purpose, the Federal Government should develop *information and discussion* materials as part of its digitalisation strategy for the current parliamentary term, with the aim of improving skills on the part of consumers, multipliers and decision-makers. The underlying principles and quality aspects of scoring, as well as forms and causes of unequal treatment are just as much part of this basic knowledge as the rights enjoyed by those scored.
2. Measures should be taken to foster the competence people require in order to take informed decisions concerning their participation in a scoring process. This includes having the skills to identify scoring services and seek alternatives, as well as to verify, assess (e.g. is the information relevant to the consumer disclosed?) and utilise such services.

3. Identifying and revealing discrimination

1. The Advisory Council for Consumer Affairs recommends that consumer information rights, as set out in Article 15 para. 1 letter h of the GDPR, be strengthened. In particular, consumers should be able to ascertain how scores are distributed among different groups with different protected attributes (to the extent that this can be established by the services themselves). This will allow consumers to provide evidence of algorithmic discrimination.
2. The Advisory Council also recommends strengthening the position of supervisory authorities (see recommendation 7).
3. Furthermore, it recommends that associations be given the right to pursue representative actions in cases of discrimination through scoring.

4. Ensuring that non-telematics based options remain available

1. The Advisory Council for Consumer Affairs recommends the introduction of legal guarantees to maintain telematics-free options for those seeking insurance (especially motor vehicle liability insurance and health insurance). In particular:
2. Policyholders who do not use telematics-based tariffs may not suffer substantial disadvantage compared to the holders of telematics-based policies.
3. Most members of the Advisory Council for Consumer Affairs believe that telematics policies should

be self-financing and should not be offered at the expense (even indirectly) of policyholders who do use telematics. Since solidarity objectives are relevant particularly in health insurance, steps would need to be taken to prohibit cheaper telematics tariffs that exist only because they attract policyholders with above-average health and do not significantly reduce the expenses incurred by insurers.

4. The use of proxy variables, as for example in geo-scoring, requires special justification (there must be a causal connection!) and must be subject to the scrutiny of the relevant supervisory authority. The use of proxy variables should be minimised. Where proxy variables are used, plausible reasoning must be given as to their substantive connection with the target variable.

5. Ensuring score quality

1. The Advisory Council for Consumer Affairs recommends that ambitious quality principles be developed on the basis of best practices. This should be based on existing quality assurance initiatives for algorithmic processes. These quality principles should be developed and updated (drafted, implemented, monitored) on a collaborative basis by industry, supervisory authorities, consumer organisations and the market watchdogs of Germany's consumer advice centres.
2. Scoring services operating in sensitive fields should be obliged to file information with supervisory authorities that is verifiable in detail and reveals the high quality of their procedures. Only then will it be possible to test scores for consumer fairness. This obligation would apply to scores which use statistical measures to predict behaviour (e.g. false positive rates, hit rate, gini coefficient, area under the ROC) for the population as a whole and for relevant population groups (by sex, age, education etc.). This would also make it possible to identify discrimination and cases of questionable score quality.
3. As the situation currently stands, scoring procedures that pursue objectives which have not been appropriately identified to the public are prohibited by law. In addition to the role of supervisory authorities (see recommendation 7), consumer organisations or the market watchdogs of Germany's consumer advice centres could also apply their expertise and contribute to uncovering "falsely labelled" scores as well.

6. Ensuring data quality

1. When developing scores, a sufficient level of data quality must be ensured and documented for supervisory authorities.
2. Scoring services and users should enter into voluntary commitments to improve their data governance, in particular their data quality management, in accordance with the standards set in the quality principles.
3. In applying the procedure, measures must be taken to ensure that data is accurate, complete and up-to-date.
4. In its report on *Digital Sovereignty*, the Advisory Council for Consumer Affairs already outlined the option of a *data dashboard*, which would allow consumers to scrutinise their own data. This would facilitate consumer-oriented data management. The Advisory Council reaffirms its recommendation that this option be explored. Such explorations should cover current developments in the area of secure identity management via blockchain-based systems, which allow consumers to manage their own identity data securely and definitively.
5. The Advisory Council recommends that research be conducted promptly to appraise and, where applicable, improve the quality of data used in relevant scoring processes, with a particular focus on entity recognition. Where necessary, improvements should be made via statutory provisions. Measures must be taken to ensure that a score calculated for

a certain person is correctly assigned to that person. The duty for providers to inform individuals that they are being scored (see recommendation for action 1) will serve to minimise the risk of identity mix-ups.

In this regard there is clearly a conflict between the interests of scoring services and users, on the one hand, and data protection interests on the other. For this reason the Advisory Council recommends that the Federal Government's Data Ethics Commission discuss ways of improving entity recognition and develop concrete recommendations.

7. Improving oversight

1. The Advisory Council for Consumer Affairs recommends that the Federal Government explore whether a *digital agency* (see the Advisory Council's report on "Consumer Law 2.0") could act as a competence centre to assist supervisory authorities in exercising their mandates. This might consist, for example, in setting up a federal institute as a centre of method expertise for quality assurance, which could also be used for "non-digital" purposes.
2. The responsible supervisory authorities should be put in the position (both structurally and in part through salary improvements for specialists, especially in statistics and IT) to perform the aforementioned tasks. Developments at the Federal Financial Supervisory Authority (BaFin) over the last few years could serve as good practice. The responsible supervisory authorities should be granted the considerable financial resources required for them to perform the aforementioned additional tasks and test concrete scoring services.
3. To ensure that the present recommendations are promptly implemented, the Advisory Council for Consumer Affairs proposes the creation of a task force at the level of the Federal Government (for example at the Federal Chancellery) in order to develop guidelines for the elaboration of quality principles on the basis of existing procedures (e. g.

at BaFin). This task force should be set up immediately after the Data Ethics Commission has finished its work.

8. Preventing "super scores"

The Advisory Council for Consumer Affairs recommends that developments in China and in other countries which are experimenting with "super scoring" are closely followed and analysed. In particular, public debate is required on the change in social values and structures that such systems entail.

The development of "super scores" by international commercial actors may also have an impact on Germany. Lawmakers and supervisory authorities should prepare for an examination of whether measures can and should be taken to ensure that "super scores" cannot be offered commercially in Germany.

The Advisory Council recommends that an examination be carried out into the extent to which existing instruments (especially purpose limitation and the "no tie-ins" rule) contained in the GDPR may also be used to prevent "super scores".

Members and staff of the SVRV

Members of the SVRV

Professor Lucia Reisch (Chair)

Professor of Intercultural Consumer Research and European Consumer Policy at Copenhagen Business School

Dr Daniela Büchel (Vice-Chair)

Member of the Trade Germany Board, REWE Group, Managing Director of REWE Markt GmbH and of Penny-Markt GmbH

Professor Gerd Gigerenzer

Director of the Harding Centre for Risk Literacy at the Max Planck Institute for Human Development in Berlin

Helga Zander-Hayat

Member of the Board of Management of North Rhine-Westphalia Consumer Advice Centre

Professor Gesche Joost

Professor of Design Research at the University of Fine Arts, Berlin

Professor Hans-Wolfgang Micklitz

Professor of Economic Law at the European University Institute in Florence

Professor Andreas Oehler

Professor of Finance at the University of Bamberg and Director of the University's Research Centre for Household Finance and Financial Literacy

Professor Kirsten Schlegel-Matthies

Professor of Home Economics at the University of Paderborn

Professor Gert G. Wagner

Max Planck Fellow at the Max Planck Institute for Human Development in Berlin, Research Associate at the Alexander von Humboldt Institute for Internet and Society, Berlin, and Senior Research Fellow for at the German Socio-Economic Panel Study at the German Institute for Economic Research (DIW Berlin)

Staff of the SVRV

Head of the Bureau:

Thomas Fischer, M.A.

Research staff of the Bureau:

Johannes Gerberding

Dr Christian Gross

Dr Ariane Keitel

Sarah Sommer, M.A.

Table of contents



About this report 13

I. Introduction	14
II. Scores and scoring	16
III. Objectives of the report	20
Objective 1: Improve the information base and increase knowledge of scoring	20
Objective 2: Broaden the empirical basis and address legal issues	21
Objective 3: Suggest rules for consumer-friendly scoring	21



Areas for action: the state of research 25

I. Transparency and comprehensibility	26
1. Transparency in predictive scoring	26
2. Transparency in behavioural scoring	27
3. Keeping transparency and comprehensibility of scoring systems on the agenda	28
4. Scoring transparency as a special form of algorithm transparency	30
5. Transparency as a condition for a social debate on scoring	32
II. Non-discrimination and equal treatment	34
1. What is discrimination?	34
2. Discrimination through scoring input	35
3. Score quality and non discrimination	36
4. Undesirable unequal scoring-based treatment beyond discrimination	39
III. Enforcement of rights	40
IV. Score quality	41
1. Quality of the algorithm underlying a score	41
2. The utility of newer and more complex algorithms	45
V. Baseline data	46
1. Accuracy, currency and completeness	46
2. Use of proxy variables	47
3. Weighting of input variables	48
VI. Competing fairness criteria	50

VII. Consumers and society: expectations, knowledge, competence and implications	52
1. Consumers' expectations and acceptance of scoring	52
2. Knowledge and competence	54
3. Social implications	57
VIII. The danger of a super score	61
1. Scoring models abroad	61
2. Data accumulation and data trading	65
3. Repersonalisation of anonymised data	68
4. Aggregation of data into a super score	69



Market survey: credit reference agencies, motor insurance telematics and health insurance policies **71**

I. Introduction and key issues	72
II. Survey design	73
1. Overview of providers	74
2. The questionnaires	75
III. Discussion of findings and highlighted consumer problems	76
1. Diffusion of scoring in the market segments under examination	76
2. Transparency	78
3. Score calculation and statistical quality	80
4. Behavioural effects	84
5. Discrimination	85
6. Aggregation of data and inclusion of new consumer attributes	87
7. Supervision	88

D

Public knowledge and acceptance of scoring 91

- I. Preliminary study, 2017 92
- II. Representative survey, 2018 93
 - 1. Analysis of the findings 94
 - 2. Multivariate regression analyses: presentation and discussion of findings 106
 - 3. Population survey findings: general summary and conclusions 109

E

The legal framework for scoring 111

- I. The basis in data privacy law 113
 - 1. Profiling (Article 4(4) GDPR) 113
 - 2. Automated individual decision-making (Article 22 GDPR) 115
 - 3. Scoring of probability values (section 31 of the Federal Data Protection Act) 118
- II. Rules for specific areas of activity 124
 - 1. The law governing standard business terms 124
 - 2. The law governing insurance contracts and insurance supervision 125
 - 3. Social insurance law and statutory health insurance 128
- III. Building blocks for a scoring regime 129
 - 1. Regulating the ‘how’ of scoring versus regulating the ‘whether’ 129
 - 2. Scoring regulation and algorithm regulation 130
 - 3. Guaranteeing a defined score quality 130
 - 4. Guaranteeing transparency and comprehensibility 132
 - 5. Guaranteeing non-discrimination 135
- IV. Supervision 138



About this report



I. Introduction

Under the heading of ‘scoring’, this report examines algorithmic decision-making processes involving direct consumer contact.¹ In so doing, it follows on directly from the discussions in previous reports from the Advisory Council for Consumer Affairs, particularly *Consumer Rights 2.0* (SVRV, 2016) and *Digital Sovereignty* (SVRV, 2017a). The subject of consumer scoring which was chosen for this report is assuming ever greater significance because of its topicality and its increasingly wide use (see, for example, Christl and Spiekermann, 2016, and Mau, 2017). In many spheres of people’s lives, increasingly complex methods are being used to analyse consumers’ characteristics and activity, predict their future behaviour or encourage them to adopt modes of behaviour that will improve their score. The product of this analysis is an individual score that can serve as a basis for establishing:

- whether and on what conditions a consumer can obtain a mortgage, for example,
- how much discount a consumer can obtain from his or her motor insurance premium for good driving, and
- whether someone is taking sufficient preventive action to qualify for a bonus from his health insurer,

and much more besides.

These are examples from three major areas of life and consumption, namely finance, mobility and health care, in which scoring is used today. These three areas have been selected for this report.

In a market economy, scoring – particularly credit scoring – plays an important role in creating transparency and trust between the two sides of the market, and, for example, new score-based insurance products certainly offer benefits for consumers. Besides such beneficial effects, however, scoring can also have unintended adverse effects.

While the SVRV is fully aware of the potential of modern scoring systems, the focus of this report is on possible risks and ways of minimising them. Our specific goal is to examine what sort of form consumer-friendly scoring – which must first be defined – might take in terms of procedure and substance, what requirements it must meet in the light of consumer policy and how such consumer-friendly scoring can be politically and institutionally underpinned. These reflections are directly relevant to the regulation of algorithmic decision-making practices in general as well as to society’s assessment and regulation of artificial intelligence² and to data ethics.³

Scoring, the formalised rating of individuals with the aid of a numerical figure, has a certain tradition in our culture; one need only think of school test and examination marks. Digitisation is now multiplying the means of rating people and, therefore, increasing the risks arising from such assessments. On the other hand, digitisation is also creating opportunities, because formalised scoring can be less discriminating as compared to informal decisions taken by individuals, such as landlords or employers. Numerous operations are embedded in any complex ‘decision-making architecture’ in which both

1 The following are examples of other algorithmic decision-making processes that are not covered by this report:

- consumer-related processes such as: 1. personalised vouchers in supermarkets and micro-targeting by online shops; these are based on information about the equipment with which users surf the Internet and about their browsing history to show personalised ads and offer personalised prices in order to induce users to make purchases, to harness customers’ propensity to spend and to ensure customer retention by means of special offers (Hosell and Schleusener, 2016; Zander-Hayat, Domurath and Gross, 2016; Zander-Hayat, Reisch and Steffen, 2016); 2. robo-advisers that assist in the selection of financial products (Oehler, Horn and Wendt, 2016); 3. algorithm-controlled self-driving cars and other largely autonomously operating products, such as cleaning robots and robotic lawnmowers.
- processes not directly relating to consumers such as: 1. people analytics (human-resources management; see, for example, Höller and Wede, 2018, written from a trade-union perspective), including applicant scoring (pre-employment screening and e-recruiting; see, for example, Christl, 2017); 2. predictive policing (see, for example, Egbert, 2018, and Sommerer, 2017).

2 On 28 June 2018, the Bundestag appointed a Study Commission on Artificial Intelligence, subtitled “Social Responsibility and Economic Potential”. The Bundestag homepage states that “The task of the Commission is to formulate practical recommendations for dealing with artificial intelligence (AI). It is to be appointed without delay and present its concluding report, including practical recommendations, after the 2020 summer recess”. German text at <https://www.bundestag.de/dokumente/textarchiv/2018/kw26-de-enquete-kommission-kuenstliche-intelligenz/560330>, accessed on 17 August 2018.

3 The Federal Government appointed a Data Ethics Commission to examine this issue. According to the homepage of the Federal Ministry of the Interior, “There is great potential in the use of algorithms, artificial intelligence and digital innovations. At the same time, they raise numerous ethical and legal questions. (...) The purpose of the Data Ethics Commission is to develop, on the basis of scientific and technical expertise, ethical guidelines for the protection of the individual, the preservation of social cohesion and the maintenance of well-being in the information age. With the Federal Ministry of the Interior, Building and Community and the Federal Ministry of Justice and Consumer Protection acting as the lead ministries, it will make practical recommendations by the summer of 2019 and propose regulatory options.” German text at <https://www.bmi.bund.de/de/themen/it-und-digitalpolitik/datenethikkommission/datenethikkommission-node.html>, accessed on 12 August 2018.

human decision-makers and machines are involved. Machines prioritise, sort and classify so as to focus the attention of human decision-makers. They stake out the area within which human autonomy of action can unfold and prestructure human decision-making processes. Humans do not normally take decisions in a vacuum, so to speak; on the contrary, their decisions add another thread to an already complex social fabric.

Before the potential of modern scoring systems can be fully exploited, a number of conditions must be met to ensure that the scoring is as consumer-friendly as possible. Cases of mistaken identity must be ruled out or at least minimised, and objections must be easy to make in practice. There must be no direct or indirect unwarranted discrimination on grounds protected by law, such as gender. In the case of consumer-friendly scoring for predictive purposes, the criteria used and the predictions themselves must be of demonstrably high quality. The predictive power of the scoring system must also be consistent across a whole range of socio-economic groups.

Not only with regard to the actual fairness of the chosen method, transparency and comprehensibility are the alpha and omega, as they are, for instance, when university admissions are allocated on the basis of school grades. Scoring systems must not mislead their subjects, for example by making hollow promises about their health or assessing people on the sole basis of their membership of a group living in a particular area. On the contrary, scores must make reliable predictions, for example about the extent of a consumer's creditworthiness. They should not, moreover, draw hasty conclusions concerning other spheres of life, as happens systematically in Chinese forms of citizen scoring. In particular, scoring must be comprehensible to its subjects, and that cannot be achieved by transparency alone. In addition, consumers should be educated about the purpose of scoring and its potential quality and discriminatory capacity, and steps should be taken to foster critical competence and so enhance the public debate.

Scoring defined

Scoring is the assignment of a numerical value (a score) to a person for the purpose of predicting or guiding that person's behaviour. That numerical value is normally determined by applying an algorithmic procedure (computer program) to a broad information basis.⁴

If these conditions, which we shall address in more detail in our recommendations, are met and if their fulfilment is systematically and effectively verified and overseen by the state, consumers need have no qualms about the further development of digital scoring processes.

On the basis of this discussion, we shall propose eight recommendations for action, the implementation of which would be desirable as part of a foresighted consumer policy in a digitised living and consumption environment. The recommendations are prefixed to the report and are explained in Part F by reference to specific problem scenarios. Our recommendations should also be worth exploring in connection with other scoring applications such as people analytics, that is to say indicator-based human-resources management.

⁴ Unlike the practice known as profiling, a process that precedes scoring and comprises the collection and correlation of data from a large number of people and identification of patterns within that body of data. One purpose of profiling, then, is to prepare for subsequent scoring. On profiling as a legal concept (Article 4(4) of the EU General Data Protection Regulation (GDPR)), see section E.I.1 below.

II. Scores and scoring

Among the best-known forms of scoring are credit scores, which are assigned to individual consumers by private credit reference agencies. Credit reference agencies collate a wide range of consumer data, such as information on a consumer's credit history, records of so-called payment irregularities and personal information about the consumer. Some credit reference agencies also include data concerning the area in which the consumer lives (geo-scoring).⁵ From this collection of data the agency derives a probability rating or behavioural prediction on the individual's creditworthiness or the likelihood that the loan will be repaid (see, for example, Schröder, Lang, Lerbs and Radev, 2014; Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). Creditworthiness is expressed in a score which can be used as an aid by banks, for instance, when deciding whether to grant loans or by online traders when determining which payment options to offer a particular customer. Depending on the score, in other words the degree of creditworthiness, of a given prospective customer, online shops offer easier – albeit riskier from the shop's point of view – payment terms, such as purchase on account, to some customers but not to others. In e-commerce, scores are used primarily for the purpose of fraud detection, to distinguish notoriously bad payers from customers with default-free records. Otherwise the latter would have to help foot the bill for losses caused by the former, which, despite fraud detection, amounted in 2017 alone to more than 2.5 billion euros, according to the German e-Commerce and Distance Selling Trade Association (bevh).⁶

In the debate on consumer policy, scoring has featured predominantly in the context of credit checking (see, for example, Bala and Schuldzinski, 2017; Oehler, 2017; Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). Against the backdrop of the progressive development of digital technology and the use of algorithmic decision-making methods in other areas of business, such as health and motor insurance, however, the focus of today's debate in

the realm of consumer policy is on novel forms of scoring.⁷ Especially as a result of progress in what is known as narrow artificial intelligence (Ashley, 2017; Goodfellow, Bengio and Courville, 2016; Jentzsch, 2018; Nilsson, 2009; Russell and Norvig, 2010; Witten, Frank, Hall and Pal, 2016) there is new scope in the field of machine learning for automated data analysis based on pattern recognition.

These new forms of scoring, the effects of which are still little known and which regulators have largely not yet analyzed, are one of the focal points of this report. One example that is examined in detail relates to the telematics-based system of pay-as-you-drive (PAYD) insurance cover, which is already an integral component of some motor insurance products. In this type of system, insurers – or, as the case may be, their contracted data-analysis agencies – record, by means of a smartphone app, for example, details of policyholders' driving behaviour, including data on journey times and routes, and communicate to the driver and his insurer a score indicating how safely the vehicle has been driven. Especially 'good' drivers, in other words those whose scores exceed a certain threshold, then receive a discount on their insurance premiums. The scores in this case serve not only to predict drivers' behaviour but can also be used by insurers for the express purpose of modifying that behaviour. At the present time there is no such thing as a pure PAYD tariff, in which the rate of premiums depends entirely on the registered score. Instead, these are always offered as an optional addition to a motor insurance policy with conventionally calculated premiums.

Other examples that are described in detail in this report are from the realm of health care. Many statutory health insurance funds reward their members with a credit note or other inducement if they collect scoring points in a bonus programme by engaging in any of a predefined set of healthy activities – preventive measures such as physical exercise, which may be recorded, for example, by a fitness tracker, inoculations and attendance at health courses (Baun and Nürnberg, 2015). According to an

⁵ According to information from the German credit bureau SCHUFA, however, this occurs only in "a few exceptional cases" in which no other information is available. SCHUFA website, <https://www.schufa.de/de/ueber-uns/daten-scoring/scoring/scoring-schufa/>

⁶ https://www.bevh.org/nc/veranstaltungen/details/datum/2017/oktober/artikel/betrug-im-e-commerce-und-datenschutz/fe_pw/?cHash=b38d2f5410c449712c2e5c2a4c0b0b1b&sword_list%5B0%5D=betrug; accessed on 3 September 2018.

⁷ Development of "the latest technological processes such as machine learning" for fraud prevention, among other things, is also benefiting credit reference agencies. Schufa website, URL <https://schufa-wegbereiter.de/de/digital/innovationen-labor/innovationen-aus-dem-labor.jsp>; accessed on 17 August 2018.

opinion delivered by the Federal Insurance Office (Bundesversicherungsamt – BVA), however, there is a need to take a close critical look at the actual beneficial health effects of the defined activities (Bundesversicherungsamt, 2018). The statutory health insurance funds are bound by the provisions of section 20(3) of Book V of the German Social Code (Sozialgesetzbuch) to operate within narrow limits when selecting eligible activities. Both private health insurers and other non-statutory insurers are freer to shape their own scoring-based prevention and fitness programmes; the Generali insurance company, for example, is planning to offer what it calls the *Vitality* Programme in 2019 within its range of private insurance products.⁸ Part of the programme is based on health scoring, whereby both participation in preventive measures and transmission of vital parameters earn points, which are redeemed with vouchers from partner companies and lower insurance premiums. Anyone who signs a non-smoker's declaration, for example, earns 4,000 points in a year; anyone buying "healthy foods" (fruit, vegetables or fish) from a cooperating online shop can earn up to four times as many points in a year.⁹

What is not clear are the criteria used to weight individual activities and the extent to which the score is actually a valid basis for conclusions regarding the improvement of a person's state of health. Generali itself sets out its stall clearly, emphasising the scientific basis of the *Vitality* Programme ("*Vitality* is a health programme based on scientific findings"¹⁰) and advertising on that basis for the health effects of participation in the *Vitality* Programme ("*Vitality* members have lower health costs"¹¹). However, impact studies showing evidence of actual improvements in people's health resulting from participation in the *Vitality* Programme in Germany in particular, including comparisons with a randomised control group, have yet to materialise, which has prompted critics to describe *Vitality* as a kind of cashback scheme using health data.¹²

Although other scoring processes in the health sector other than the *Vitality* Programme are still few and far between, growing public acceptance of self-measurement, including by means of wearable devices, which is discussed in *The Quantified Self* and *Self-Tracking* (see, for example, Lupton, 2016, and Selke, 2014) seems to indicate that this could change in future. Providers are already cottoning on to the fact that many people now record and analyse their physical performance data for the purpose of improving their fitness levels. This practice is encouraged by statements such as the one made by the start-up business Dacadoo to the effect that a person's health is improved by the use of apps which convey the user's state of health in the form of a numerical value.¹³ This development is part of a general tendency in preventive health care to focus increasingly on early detection and preventive health-promoting action as a supplement to curative treatment (see, for example, GKV-Spitzenverband, 2017). A study commissioned by the Federal Ministry of Health found that there was still a lack of evidence in the form of robust studies which would allow a conclusive identification of beneficial health effects of fitness apps, particularly of the longevity of any such effects (Albrecht, 2016).

Regulatory measures such as the e-Health Act of 1 January 2016, designed to establish a modern IT infrastructure in the health sector, and the loosening, which took effect on 10 May 2018, of the ban on remote treatment to allow the practice of telemedicine, for example in video consultations, are also indicative of a general increase in the use of digital services in the health sector (see also Gigerenzer, Schlegel-Matthies and Wagner, 2016). On the one hand, this is a gain for patients such as those living in rural areas with few doctors and even fewer specialists; on the other hand, these technological solutions create a data problem on an unprecedented scale.

⁸ Because of regulatory requirements, these are confined for the time being to the fields of disability and occupational disability insurance and term life insurance.

⁹ URL <https://static01.cosmosdirekt.de/CosmosCAE/S/linkableblob/home/213750.1525232169000/data/Antragsinformationen- zur-Generali-Vitality-Mitgliedschaft-data.pdf>; accessed on 4 September 2018.

¹⁰ URL https://www.generallivitality.de/vmp/so_funktioniert_vitality; accessed on 4 September 2018.

¹¹ German text at URL https://www.generallivitality.de/vmp/so_funktioniert_vitality; accessed on 4 September 2018.

¹² URL <https://www.sueddeutsche.de/wirtschaft/versicherung-wer-sich-bewegt-zahlt-weniger-1.2920176>; accessed on 4 September 2018.

¹³ URL <https://info.dacadoo.com/de/unternehmenslosungen/life-health-insurance-solutions/>; accessed on 19 June 2018.

A score that provides information on a person's own state of health can certainly be regarded as a means of consumer empowerment, as it reduces the information asymmetry between doctors and patients. It is questionable, however, whether the same applies to the information asymmetry between consumers and companies; on this point, the German Ethics Council takes the view that such asymmetry is more likely to be widened by the use of big data, which potentially enable companies to find out more about their customers (German Ethics Council, 2017; cf. Weichert, 2018). Both research and implementation, however, are still at quite an early stage, and besides highlighting the opportunities, it is worth sounding an early warning of the risks which may not surface until later and which can easily be overlooked by consumers because of the immediate benefits.

In contrast to the discussion of these relatively new applications of scoring, the debate in the field of consumer and market policy on the role of credit scoring in the financial sector has been going on for many decades. There is generally a good stock of literature, on the basis of which the macrosocial advantages of credit scoring may be summed up as follows: the use of credit scores reduces loan defaults; it lowers transaction costs and therefore has a major impact on the efficiency of financial markets (Schröder *et al.*, 2014). In addition, credit scores can help to reduce information asymmetries that exist between borrowers and lenders and to prevent credit rationing, because they give lenders the vital information they need about prospective borrowers when it comes to granting loans (Schröder *et al.*, 2014). In the realm of online shopping, credit scores play a major part in the detection and prevention of online fraud (Bolton and Hand, 2002; Marschall, Morawitzky, Reutter, Schwartz and Baars, 2015). On the other hand, there are legitimate concerns about data sovereignty, for example, or about discrimination against particular groups, as we explained in previous reports, such as SVRV, 2017a.

A similar situation exists with regard to the new telematics-based tariffs for motor insurance. On the one hand, they can lead to greater safety and better traffic flow, less information asymmetry and more efficient markets. Insurers advertise that continuous recording and analysis of speed and acceleration data encourage careful driving, thereby contributing to greater road safety.¹⁴ The analysis of individuals' driving behaviour can also serve as a basis for more risk-related rewards and discounts (Baecke and Bocca, 2017; Bian, Yang, Zhao and Liang, 2018), which can be especially beneficial to young drivers, who are otherwise charged very high premiums. On the whole, so the argument goes, driving analysis allows a more precise actuarial cost calculation for motor insurance (see, for example, Baecke and Bocca, 2017; Bitkom, 2014; Kraft and Hering, 2017). Another socially desirable potential benefit of more careful driving as a result of scoring consists in a reduction in congestion and environmental pollution (Kraft and Hering, 2017; Litman, 2005).

Concerns are expressed to the effect that constant recording and analysis of driving behaviour can lead to increasing surveillance by commercial insurance firms (see, for example, Stiftung Warentest, 2014, and Verbraucherzentrale Bayern, 2016). Last but not least, the criticism is quite often made that, while telematics-based deals benefit consumers, their main beneficiaries are insurers themselves, which exploit the increased opportunities to address consumers directly, through push notifications on smartphones for instance, as a means of customer retention (see, for example, der Weidner and Transchel, 2015).

To put it plainly, this report does not contest the fact that scoring performs an important function in business and society. The real question for the SVRV is how scoring is and should be designed. Scoring-based business models are normal today,¹⁵ even though they are applied in varying depth, and can bring many benefits for individual consumers and for markets in general. They also entail risks, of course, some of which are already obvious, while others are only just beginning to emerge and, given the rapid speed of technological development, cannot by any means be definitively assessed.

¹⁴ <https://www.cosmosdirekt.de/veroeffentlichungen/versicherungstipp-telematik-198254/>; accessed on 19 June 2018.

¹⁵ See also footnote 1, which refers to applications of scoring that are not dealt with in detail in the report, namely micro-targeting by online shops, robo-advisers helping with the selection of financial products, applicant scoring, predictive policing and, in particular, the individualised control of social media by their providers.

This report highlights the key challenges of scoring-based business models and makes recommendations on political measures that can strengthen the position of consumers. At the heart of the report is the concept of consumer-friendly scoring, and our task is to describe that concept (see also Mittelstadt, Allo, Taddeo, Wachter and Floridi, 2016) and to discuss the following questions:

- What does fairness mean in the context of scoring-based business models?
- Which data should be included in scores, and which should be excluded?
- Which statistical quality criteria should scores meet?
- Which assessment criteria are relevant to consumer-friendly scoring?
- What does discrimination through scoring mean, and where and how does it occur?
- Which elements of scores should be known, which should be made transparent and comprehensible, and which should not?
- Which forms of transparency and monitoring should there be to ensure that scoring processes maintain or improve the enforcement of consumer interests? Are the existing processes adequate?
- Which institutions lend themselves to the tasks of creating transparency and monitoring?

In a longer-term perspective, this report also considers the development of so-called super scores, in other words scores that not only assess consumer behaviour within a limited area such as finance, mobility or health but assess it across the board. We shall look closely at Chinese pilot projects for a system of social credits in which, from 2020 onwards, all citizens of the People's Republic are to be assigned an individual score that will take account of behaviour patterns in all areas of their lives (Kostka, 2018). In view of the differing political and legal systems, this model is not transferable to the Western world at the present time, nor will it be for the foreseeable future, but it nevertheless provides food for thought on what is technically feasible and what is socially acceptable and unacceptable. Germany's President Frank-Walter Steinmeier spoke in similar terms on 15 February 2018: "There is no threat of such a thing [as the Chinese system] happening in Germany, but it goes to show how important it is that we engage in detailed discussion on the social implications of technological developments."¹⁶

¹⁶ German text at <http://www.bundespraesident.de/SharedDocs/Reden/DE/Frank-Walter-Steinmeier/Reden/2018/02/180215-Leopoldina-Sachsen-Anhalt.html>; accessed on 1 October 2018.

III. Objectives of the report

The particular relevance of the subject of scoring to consumers, society and business is due to three main factors: the increasing availability of individualised consumer data, the spread of methods that can be used by businesses to process these accumulations of data and to profile data subjects and the resultant growing number of applications for scoring. Novel scoring methods, moreover, no longer serve merely as a predictive tool but are increasingly being used to guide consumer behaviour too. In the present report we are pursuing three specific core objectives:

Objective 1: Improve the information base and increase knowledge of scoring

A new feature of the information base is that searches for consumer data are no longer confined to simple facts such as age and current loan agreements but can target far more detailed variables, such as vital parameters and driving behaviour, with the aid of new methods. At the same time, inexpensive means of data storage are constantly developing. The growth of business models for consumer-data brokerage by companies such as acxiom and Oracle, moreover, indicates there is already, in principle, sufficient business interest and know-how to drive the compilation of extensive data sets on consumers – if that became legally permissible – which might enable companies to develop far more complex scores that were no longer be confined to a single area of activity, such as finance, but covered many aspects of people's lives. In this way they would no longer merely lay the foundations for – or even directly make – decisions relating to people as consumers in a demarcated sector but would determine the individual's stake in the economy and society in general.

There are many processes that permit an automated assessment of whether, for instance, a person represents a high credit risk or is a good driver with a low accident risk: these range from simple rules of thumb – also known as heuristic approaches – through standard statistical estimation methods such as logistic regression, which is used in credit-scoring practice and constitutes a fairly simple form of machine learning, to new deep-learning methods, such as those based on neural networks, which process patterns and correlative connections between numerous variables (e.g. consumer attributes) based on large data sets in largely automated operations.

From the perspective of consumer policy, it is clear that the less transparent a process is, the harder it is for supervisory authorities to oversee, which also jeopardises its comprehensibility to consumers. Whether the use of machine learning will inevitably reduce the fairness of scores or could increase it is the subject of intense discussion.

In this report we not only discuss the algorithms that are currently used in consumer scoring but also focus particularly on individual and institutional means of reviewing and regulating these algorithms. Special emphasis is laid on the establishment of minimum quality standards for model scoring methods and on issues of comprehensibility for consumers.

On the whole, the increasing availability of data and the spread of new methods are making it possible to develop novel applications of scoring processes that go far beyond traditional credit scoring. This report examines these by reference to examples from the fields of mobility (telematics in motor insurance) and health (bonus programmes of statutory health-insurance funds, first steps towards health scores on the part of private health insurers and health scores already used by companies such as Dacadoo). Whereas traditional credit scoring is limited to the prediction of future behaviour, novel scoring schemes that analyse exercise habits and, in some cases, vital parameters and communicate the resulting score 'live' to the consumer are much more likely to have a behaviour-modifying effect too. The question whether this is desirable for society as a whole must be discussed.

Objective 2: Broaden the empirical basis and address legal issues

With this report, the SVRV is fulfilling its mission of creating a broad empirical basis for a well-informed and forward-looking consumer policy. The report focuses on the dynamic area of novel forms of scoring.

Following an outline of the current state of research in Part B, the SVRV presents a comprehensive study of the market segments of credit checking and mobility in Part C. In the first of these market segments, scoring methods are already long-established, while in the other they have been gaining an increasingly firm foothold in recent years. This study also examines the extent to which scoring methods have already taken root in the sphere of statutory and private health insurance. As this has only happened to a limited extent, we could use the term ‘proto-scoring’ in this context. A total of three questionnaires were devised for the purposes of this study and were sent to all identified credit reference agencies, motor insurers, statutory health-insurance funds and private health insurers. The questionnaires, which were completed in written form, were digitised, and the replies were coded for comparability and aggregation and analysed. The main purposes of this study were to examine the penetration of these three market segments by scoring practices and also to investigate which consumer characteristics were normally recorded and used to calculate the individual scores and which quality criteria were met by the algorithm behind the scoring. To add more depth to the discussion, background talks were conducted with individual companies and experts from the three sectors.

In addition, the SVRV devised a public survey and had it conducted to find out more about awareness and acceptance among the German population of established scoring practices as well as those that are technically feasible in principle but not yet established; this is described in Part D. In cooperation with a social and market research company, the infas Institute for Applied Social Sciences, a representative survey was conducted by means of the CATI (computer-assisted telephone interview) method. The sample comprised 2,215 respond-

ents. The data were subsequently processed and analysed by the SVRV.

Besides these empirical studies, a legal study, described in Part E, addressed the data-protection issues in detail, examined the current rules governing scoring and similar practices in the market segments of credit, motor insurance and health and added a set of draft building blocks for legal provisions regulating scoring in general. Part E concludes with reflections on enforceability and oversight.

Objective 3: Suggest rules for consumer-friendly scoring

In particular, the SVRV wishes to use this report to propose criteria for consumer-friendly scoring as a basis for discussion. In the view of the SVRV, scoring is consumer-friendly if scores are presented comprehensibly to consumers, if awareness of scoring and scoring skills are sufficiently available, if discriminatory elements are probed and revealed, if telematics-free options are available and will remain so in the future without significant disadvantages, if the quality of scores and data is guaranteed, if supervision of scoring is significantly improved and if the use of super scores is effectively prevented.

The history of scoring

In historical terms, the desire to assess individuals' characteristics, behaviour and preferences as precisely as possible and to infer future developments from that information is nothing new. Even in the analogue world, conclusions were – and still are – drawn about individuals from certain characteristics and modes of behaviour, and in some cases numerical values are assigned to people on that basis. Digitisation, involving complex algorithms and broad information bases, has merely injected fresh impetus into an old practice, and some believe that artificial intelligence (AI), with its auto-adaptive algorithms, has added another new dimension.

The assessment of performance and abilities by means of numerical values or standardised verbal formulas has a long tradition. In German schools, for example, marks have been awarded since the 16th century (Lintorf, 2012) and not only for learning achievements but also – as is now planned for all citizens in China (Kostka, 2018) – for social behaviour. Today, assessments of individuals' attainments and learning successes still have consequences such as admission to a higher class or a particular type of secondary school or the award of a final certificate, such as the German *Abitur*. A pupil's mark in that final examination is one of the key criteria for university admission. The *Abitur* grade is calculated as an average of the pupil's grades in all subjects, although some grades, for example those obtained in the pupil's main subjects, are weighted more heavily than others. In the admission procedure, the applicants with the best *Abitur* grades – usually in combination with their previous waiting time for a university place – are accepted until the number of available places is exhausted (the system of *numerus clausus*). In this context it becomes particularly clear that the *Abitur* grade not only represents an assessment of past performance but is also credited with predictive power as an indicator of future attainment. A good *Abitur* grade is supposed to show that the student may be expected to perform well and is likely to obtain a degree.

Another area that traditionally depends on the measurement of human performance is competitive sport. Not only performances are measured, however; in many disciplines people themselves are measured too, so that they can be categorised for the sake of creating

the most exciting possible spectacle and the fairest possible match. In boxing and weightlifting, for instance, competitors are assigned to particular weight divisions, golfers receive handicaps, and tennis players, for example, are seeded. Lastly, sportsmen and sportswomen are characterised time and again by their scores. One example is Armin Hary¹⁷, the first athlete to run 100 metres in 10.0 seconds. The unofficial electronic measure of his sprint – which was more precise than the official hand-held stopwatch – showed that he had taken about 10.2 seconds with a borderline tailwind, and his case may therefore be used to illustrate the measurement problems that can arise with all sorts of scores.

Particularly in business deals, in which contracts are very frequently concluded with hitherto unknown partners and a certain leap of faith must be made, risk minimisation plays a major role, which is why great importance has attached to scoring in this domain for many decades. Businesses need to inform themselves about the reliability and creditworthiness of customers, and in the 19th century this need led to the emergence of the first credit reference agencies. Among the first such agencies in Europe were Wys Muller, founded in 1861, Schimmelpfeng, founded in 1872, and Creditreform, founded in 1879. These agencies collected economically relevant information on individuals and companies and sold it to businesses and banks. Since then, credit reference agencies have been a cornerstone of every functioning credit system.

The first attempts to quantify people's default risk and present it as a numerical value were made in the 1940s. Until then, rudimentary scoring systems, operated by mail-order firms for example, comprised a catalogue of criteria with the aid of which sellers would verify fulfilment of a number of conditions and tally the number of ticked boxes (Thomas, Crook and Edelman, 2017). In a research project in 1941, mathematician David Durand became the first to use discriminant analysis to determine the default risk of loans (Durand, 1941). He analysed data sets on previously granted loans to identify the decisive factors that had led to smooth repayment and those that had been responsible for repayment difficulties and developed a credit score. The first firm to develop statistical models for granting credit on a com-

¹⁷ Abgerufen am 1. Oktober 2018 von URL https://de.wikipedia.org/wiki/Armin_Hary.

mercial basis was Fair, Isaac and Company, now known as FICO, in California. From the 1950s it sold scoring products to financial institutions, retailers and mail-order firms (Dixon and Gellmann, 2014).

In subsequent decades, mathematical advances were accompanied by innovations in electronic data processing, which ultimately paved the way for largely automated credit scoring. The combination of computers and algorithms as well as the experience of businesses that had seen a sharp reduction in default rates for their loans and in frauds led to the scoring products of credit reference agencies with which we are familiar today.

Another area in which such forms of risk assessment have long been established is that of insurance, where the primary purpose of scoring was, and still is, to calculate sums assured and contribution rates for each individual customer. Back in the 1920s and 1930s, German health insurers became interested in putting the calculation of health insurance contributions on a sound mathematical and statistical footing. Using what are called morbidity tables, insurers found that medical costs could be expected to vary depending on a person's sex, age and occupation (Wagner-Braun, 2002). Even today, premiums for private health policies are calculated individually on enrolment. The same applies to term life insurance and to occupational disability insurance. Consumers are categorised on the basis of a combination of individual characteristics, such as age and medical history, the risk to the insurer is assessed on that basis, and the premiums are calculated accordingly.

The calculation of premiums is normally particularly complex in the case of motor insurance, where tariffs are tailored to individual customers on the basis of numerous criteria. Among the key factors are the vehicle model type, regional weighting and the driver's no-claims history as well as characteristics such as the number of drivers, the drivers' ages, the age of the vehicle, its mileage and where it is kept (Gesamtverband der Deutschen Versicherungs-wirtschaft e.V., 2016). And there is yet another scoring system for drivers, namely the driver fitness assessment system administered by the Federal Motor Transport Authority (*Kraftfahrt-Bundesamt*) in Flensburg, commonly known as the Flensburg points system. Since 1974, the Authority has been entering penalty points in a register for administrative and criminal traffic offend-

es. When a particular number of points is amassed, the Authority issues warnings, orders drivers to attend driver fitness seminars or withdraws their driving licence (Kraftfahrt-Bundesamt, 2017).

The forms of scoring described above have a long history and already existed, to be sure, in the analogue world. Yet it is also undeniable that the way in which scoring is carried out has changed radically with the technological developments of the digital age. In France, for instance, the allocation of university places has been regulated since 2018 by a scoring algorithm known as *Parcoursup*, which analyses the applicant's fulfilment of the entrance requirements, place of residence and preferences (Joreres, 2018). In the realm of e-commerce, a consumer's creditworthiness can be calculated automatically in a matter of seconds and appropriate payment options offered. And in motor insurance, we have seen the advent of telematics-based tariffs, in which driving behaviour is constantly evaluated and scored and premiums are adapted accordingly.

Algorithm-based scoring, moreover, is being used increasingly in many new areas, assessing consumers and groups of consumers in the widest variety of ways and with the most diverse consequences (Dixon and Gellmann, 2014). There are scores that predict households' purchasing power or propensity to spend (Equifax, 2018, and Blackbaud, 2014), scores that indicate whether consumers will switch their custom to other companies (Versium Analytics, Inc., 2018), scores designed to detect pregnancies (Duhigg, 2012) and scores that measure energy consumption behaviour (Trove, 2018). Dating services are based on scores which quantify how closely personal profiles are matched (Carr, 2016).

A culture of assessment and quantification is developing (Mau, 2017). Whether it is Likes on Facebook, the number of followers on Twitter or stars on Airbnb, scoring is no longer the preserve of companies who assess consumers and assign them numerical values – it has become an everyday activity.



Areas for action: the state of research



I. Transparency and comprehensibility

What the appropriate level of transparency for scoring systems is and how that level can be achieved are unanswered questions in discussions among the general public and researchers. In the context of scoring, transparency means the disclosure of information to consumers by producers or users of scoring systems. Reflections on the right level of transparency always go hand in hand with the question how information that is made transparent should be processed and structured to ensure that it is actually comprehensible. On the one hand, the yardstick may be set for consumers, enabling them to play an informed part in scoring processes. On the other hand, the comprehensibility gauge may be set for experts to enable them to engage in critical examination of scoring systems.

Credit scoring has hitherto been at the heart of the transparency debate, because until very recently it was the most technically developed and most widespread form of scoring (on its history, see Beckhusen, 2004). Future discussions on the transparency of scoring may be expected to merge with the vigorously conducted debate of more recent times on the transparency of algorithmic decision-making procedures. The transparency aspect of scoring will remain relevant, not least because only an adequate level of transparency will enable consumers to assert more extensive rights, such as the right to correction of an erroneous score. Transparency is ultimately a prerequisite for any informed debate within society on the phenomenon of scoring.

ing interest groups, whose arguments cannot simply be dismissed and are constitutionally underpinned in each case.

An individual whose behaviour is the subject of a predictive assessment based on scoring, whom we shall refer to below as the ‘scored person’, will normally have an interest in learning that a scoring process is taking place at all. Secondly, he or she will want to know about the consequences of the resulting score. The individual may also be interested in knowing the data on which the calculation of his or her score is based, in other words which personal characteristics are taken into account in determining the score. Lastly, the individual may be interested in gaining some insight into the internal workings of the scoring algorithm, particularly the relative weight attached to each personal characteristic in the calculation of the score.

These interests may, on the other hand, conflict with the confidentiality interests of the party conducting the scoring process, referred to below as the ‘scorer’, or of the public. A scorer will generally have an interest in confidentiality if the predictive product of the scoring process is economically valuable and therefore merits protection as a trade secret; this interest was recognised by the Federal Court of Justice in its judgment of 28 January 2014 in case VI ZR 156/13, recorded in the Civil Decisions of the Federal Court of Justice (BGHZ), Vol. 200, p. 38; see also section C.III.3 below). If the details of the algorithmic method for calculating scores become common knowledge, the method ceases to be a trade secret and can be copied by competitors.

A public interest in confidentiality can exist if disclosure of the scoring method would lessen its predictive value in certain circumstances, which we shall shortly examine. This may be socially undesirable. It cannot be denied, for example, that there is a general public interest in reliable credit assessments.

Not every disclosure entails a risk of diminished predictive quality. Disclosure is harmless if the score is based on characteristics which are actually responsible for the assessed probability. In this case the scored person, by modifying his or her behaviour in a way that should serve to improve the score, is actually influencing the probability of the predicted event. Those who

1. Transparency in predictive scoring

Scoring processes are used both to predict and to modify modes of behaviour. When it comes to ensuring transparency, a distinction must be made between these two purposes. The fact is that scoring systems designed to predict behaviour are not normally meant to operate reflexively, in other words they themselves are not intended to influence the observed behaviour. Accordingly, what constitutes an appropriate level of transparency is a bone of contention between oppos-

take regular exercise reduce their risk of illness – for this reason, a person's decision to engage regularly in sporting activity cannot be described as a 'manipulation' of his or her score.

By contrast, the predictive value of the score decreases if the behaviour modification relates to variables which, though they have been good indicators of scored probability the past, do not influence the probability rating; practical examples are cited in section B.VIII.1 below. No one reduces his or her risk of illness by buying sports gear but not using it; anyone who knows that the purchase of trainers is included in a health score as a so-called proxy variable (see section B.V.2 below for more details) might thus be tempted to affect his or her score by means of consumption decisions rather than actual sporting activity. If the workings of a scoring system are revealed, scored people can recognise the effects of their behaviour on their score and therefore modify their behaviour to suit their score (Bambaucher and Zarsky, 2018).

Influencing scores by targeting non-causal criteria is discussed in literature under the heading of 'gaming the system' (Rona-Tas and Hiss, 2011). British economist Charles Goodhard encapsulated this insight into the self-reflecting nature of social systems pithily in the statement "When a measure becomes a target, it ceases to be a good measure". It merits consideration in any discussion on indicator-based control (Strathern, 1997; Wagner, 2018; Weingart and Wagner, 2015). The prevention of system-gaming may be in the general interest as a means of preserving the validity of the predictive score. A certain lack of transparency about the operation of the scoring method is then required. On the other hand, the precise opposite conclusion may also be drawn from this scope for 'gaming the system', namely that the right way to remedy the potential for manipulation is not to maintain opacity about the scoring criteria but to exclude non-causal criteria from the scoring process. This approach may be harder to achieve, but the greater fairness of a system based on causal criteria alone cannot be refuted out of hand (a detailed treatment is to be found in Britz, 2008).

2. Transparency in behavioural scoring

Scoring can also be an instrument of behaviour modification. When that is the underlying purpose, transparency seems at first sight to be an essential condition for the effective use of scoring, for an incentive system cannot achieve a targeted behavioural effect unless it reveals the connection between behaviour and its assessment. To put this in the context of the scoring process, if the scorer's aim is to motivate scored individuals to improve their score, it seems imperative that the scorer must at least disclose that certain modes of behaviour 'win points'.

However, there is also a 'softer' means of modifying behaviour through scoring. This can be illustrated by means of a hypothetical example. Imagine that a score was calculated for healthy living or good driving, but the scoring criteria were not disclosed. One might assume that such a scoring system would have effects on the behaviour of scored persons, who would try to improve their score. Only the direction of the behaviour modification in this case would be more uncertain, as the scored person can only presume what modes of behaviour are assessed by the scorer as healthy living or good driving and so count towards a better score. The scored person is therefore faced with the challenge of satisfying scoring criteria that he or she does not know.

3. Keeping transparency and comprehensibility of scoring systems on the agenda

Transparency is a key instrument of consumer policy. Accordingly, numerous studies are devoted to the legitimacy, effectiveness and limits of the transparency principle in the realm of consumer protection; a summary of this issue can be found in Tamm, 2011, especially on pages 347ff. The link between transparency on the one hand and how informed consumers actually are on the other is being increasingly questioned (Ben-Shahar and Schneider, 2014; Kettner, Thorun and Vetter, 2018; see also section B.VIII.2 below). Maximum transparency by no means implies maximum protection of consumers. Safeguarding real consumer autonomy is therefore set to move to the heart of the discussion. The debate on ‘algorithmic transparency’ could act as a catalyst, because the ineffectiveness of obligations designed only to ensure the disclosure of unprocessed information is especially evident in this context. It would be expecting too much of any consumers to present them with bare programming codes (see section 4 below).

Nevertheless, the very fact that scoring is a data-processing operation makes it reliant on a certain degree of transparency, because only a transparent system allows individuals to exercise their right to protection of their personal data (Bull, 2011). “The legality of decisions can only be verified by those who know – and understand – the data basis, the processing sequence and the weighting of the decision-making criteria” (Martini, 2017, p. 1018). This applies especially to the accuracy of the individual items of data that are used to calculate scores. Rights to rectification of inaccurate personal data (see in particular Article 16 of the General Data Protection Regulation) become irrelevant if the data subject is unaware of the inaccuracy. On the subject of actual awareness of information rights, however, see section B.VII.2 below.

The means whereby transparency is supposed to be established in the realm of scoring are legal in nature. This is why the main focal point of the academic discussion

on the appropriate level of transparency for scoring has been the dialogue between the legislative and judiciary on the one hand and legal scholars on the other. Both legislators and legal scholars have seen a particular need for regulation of credit scoring. Three events have structured the transparency debate in that area.

The first caesura came with the creation of scoring-specific data protection provisions in 2009 (Federal Law Gazette I, p. 2254). By revising the Federal Data Protection Act, the legislature assembled a body of provisions governing scoring from section 28b of the old version of the Act, which set out the requirements for lawful scoring, and from section 28a of the old version, which regulated the transfer of data to credit reference agencies. These rules were supplemented by a scoring-specific extension of the information rights of data subjects that had been enshrined in section 34 of the old version (Heinemann and Wässle, 2010). Before this new set of rules was enacted, the permissibility of scoring had been determined on the basis of the general data protection provisions. The result was not only a considerable degree of legal uncertainty as regards the very legality of scoring (Petri, 2003, and Beckhusen, 2004), reflected for example in the sceptical appraisal by the Federal Data Protection Commissioner of the time (BfDI, 1996, point 31.2.3), but also in widespread expressions of dissatisfaction with the insufficient transparency of scoring systems and with the way in which they worked in practice (Korczak and Wilken, 2008). Although some extensive information rights for data subjects have been derived from the general data-protection provisions (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, 2005), the prevailing diagnosis pointed to an inherent transparency deficit in the legal provisions (Kloepfer and Kutzschbach, 1998; Möller and Florax, 2003; Petri, 2003; Beckhusen, 2005). The legislature sought to remedy the criticised transparency shortfall by creating special scoring-specific provisions (see the explanatory memorandum for the pertinent instrument, the Federal Data Protection Amendment Bill, in Bundestag printed paper 16/10529, p. 6 *et al.*, the report and recommendation for a decision from the Committee on Internal Affairs – Bundestag printed paper 16/13219, pp. 1–2 and 10 – and presentations of the legislative project from a stakeholders’ perspective (Piltz and Holländer, 2008, and Metz, 2009)). The Amendment Act altered the basic legal framework, and so we cannot simply carry on

from the lively discussion on scoring and the identified transparency deficit that was being conducted before the adoption of the Act, because questions that were unanswered then have now been resolved by means of binding legislative provisions.

A second caesura was marked by the ‘Schufa judgment’ of the Federal Court of Justice (Federal Court of Justice judgment of 28 January 2014, Case No VI ZR 156/13, Civil Decisions of the Federal Court of Justice (BGHZ), Vol. 200, p.38). In that judgment the court clarified the scope of the information right enshrined in the first sentence of section 34(4) of the old version of the Federal Data Protection Act. The court ruled that information was to be provided on the types of personal data relating to the data subject which were used in the calculation of the score. There was no obligation, it said, to disclose information on the method used to obtain a specific score from that set of personal data and from other data. In particular, the way in which the data were weighted was not covered by the information right. As a trade secret, the scoring method enjoyed the protection afforded to fundamental rights. The judgment of the Federal Court of Justice generated keen interest among legal scholars, and the legal database *Juris* contains more than a dozen academic analyses of the decision. They form a heterogeneous picture, ranging from emphatic endorsement (Taeger, 2014) to criticism (Gärtner, 2014, and Schulte am Hülse and Timm, 2014). From now on the judgment would be the main reference point of the transparency decision. The authors of the evaluation report on the new scoring provisions of 2009 for the Federal Ministry of Food, Agriculture and Consumer Protection and later for the Federal Ministry of Justice and Consumer Protection subjected the judgment to detailed analysis and criticism (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). The plaintiff in the Schufa case has lodged a constitutional complaint with a view to overturning the judgment. The Federal Constitutional Court has not yet ruled on her complaint. A legislative initiative from the opposition ranks (Bundestag printed paper 18/4864) sought to alter the legal position but was ultimately unsuccessful. The purpose of the bill was the enshrinement in the Federal Data Protection Act of more stringent transparency requirements than the Federal Court of Justice had imputed from the Act in its Schufa

judgment. The duty of disclosure was to be extended to include “the utilised items of data, the weighting of the utilised data, the utilised comparison groups and the assignment of the persons concerned to the comparison groups whose data are used in the calculation of the probability value” (*ibid.*, p. 4).

The entry into force of the General Data Protection Regulation (GDPR) in May 2018 marks the third caesura in the transparency debate. The GDPR replaced a system of national data privacy laws under an umbrella of EU law with a directly applicable European legal instrument. The Regulation diverges in many respects from previous data privacy law, and not only in its material scope; it does not forge an unbroken link with the established terminology, regulatory method and legislative style of the old Federal Data Protection Act either. It therefore seems plausible that the adoption of the General Data Protection Regulation marks the “start of a new era in data privacy law” (Schantz, 2016). The GDPR is peppered with flexibility clauses that give national legislators discretionary powers. On this basis a new Federal Data Protection Act was enacted to supplement the General Data Protection Regulation. Section 31 of the new Act contains a special scoring-specific provision (for more details see section E.I.3 below, which also examines the conformity of the provision with EU law. With this provision, headed “Protection of commercial transactions in the case of scoring and credit reports” the German legislature sought “to preserve the material protective standard of sections 28a and 28b of the Federal Data Protection Act, old version”, as the explanatory memorandum to the new Act puts it (Bundestag printed paper 18/11325, p. 101, which corresponds to Bundesrat printed paper 110/17, p. 101). Initial academic studies on scoring under the General Data Protection Regulation do not expect the new regulatory regime to bring radical changes (Taeger, 2016; von Lewinski and Pohl, 2018) and indeed the resilience of ingrained practice in the face of legislative innovations must not be underestimated. Nevertheless, the legislative design of the transparency requirements and information rights in Articles 13, 14 and 15 of the GDPR differ sharply from sections 19 and 34 of the Federal Data Protection Act in its old version (for more information, see section E. III.4 below).

The transparency of scoring methods is not only discussed in academic circles but is also a subject of public debate. In February, the non-profit organisations Open Knowledge Foundation and AlgorithmWatch launched the OpenSchufa initiative. One of the declared aims of the project is to ‘crack’ the algorithm with which Schufa obtains its credit scores (OpenSchufa, 2018). The plan is to find out both the data that go into the calculation of the score and the method by which the score is obtained from that information material by asking as many people as possible to reveal their Schufa score and their personal details. Numerous media have reported on the aim of the initiative; examples are given in Erdmann, 2018, and Schneider, 2018; Schufa itself responded critically to it (Schufa Holding AG, 2018a).

4. Scoring transparency as a special form of algorithm transparency

There is currently a lively debate on a suitable regulatory regime for digital algorithms. This debate is not only taking place in academic circles but is also commanding the close attention of German politicians. In the coalition agreement between the CDU, CSU and SPD government fractions for the 19th legislative term, regulatory goals were formulated for algorithmic decisions (CDU/CSU/SPD, 2018, lines 2092ff.). Policymakers in the field of consumer affairs (CDU/ CSU/SPD, 2018, lines 6266ff.), associations (Gesamtverband der deutschen Versicherungs-wirtschaft e.V. (German Insurance Association), 2018, and Verbraucherzentrale Bundesverband e.V. (Federation of German Consumer Organisations), 2017) and bodies from civil society (see, for example, the initiatives presented at www.algorithmenethik.de and at www.algorithmwatch.de) have also been discovering the subject of algorithmic transparency for them-

selves. The SVRV has set out its basic position (SVRV, 2016; SVRV, 2017; SVRV, 2017a; summarised in Micklitz, 2017), stressing the need to ensure, by means of legal prescripts, that the underlying parameters of algorithms relating directly to consumers are made transparent and disclosed in a standardised format to a group of experts from a regulation agency for digital operations¹⁸ (SVRV, 2017a; more in Gigerenzer, Wagner and Müller, 2018).

A scoring algorithm is one particular type of algorithm (Just and Latzer, 2016). The discussion on scoring transparency can therefore be conducted as part of the general debate on the regulation of algorithms. The conventional scoring methods of the present time are significantly less complex than the algorithmic decision-making systems which usually serve as reference points in the debate on algorithm regulation (SVRV, 2016) and which are not infrequently assignable to the realm of artificial intelligence. However, even the algorithms that are used today are not easily understood by non-experts (see section B. IV.2 below for more details). And if the complexity of practised scoring methods were to increase, for instance in the direction of methods based on systems of machine learning, particularly neural networks (see Hurley and Adebayo, 2016; Thomas, Crook and Edelmann, 2017), the debate on algorithmic transparency would also become increasingly relevant to consumer scoring. Whether this increasing complexity of scoring methods will actually materialise on a wide scale, making scoring systems into ‘black boxes’, is uncertain, not least for the simple reason that it has yet to be determined whether such new scoring systems are sufficiently superior to the conventional methods to make their use economically justifiable. So there is no evidence yet that novel algorithmic decision-making methods are always ‘better’ – in terms of model accuracy, for instance – than established methods. Pertinent examples are Google Flu Trends, designed to predict flu epidemics, and COMPAS, designed to assess the likelihood that an offender will re-offend; in both cases, the predictive capacity of complex algorithms has been found inferior to that of simple rules of thumb (Dressel, 2018; Lazer, Kennedy, King and Vespignani, 2014).¹⁹

¹⁸ On such an agency, see also Tutt (2017), who advocates a central regulatory authority for algorithms modelled on the Food and Drug Administration and outlines the powers of such a body (pp. 105ff.): “The agency should serve as a centralized expert regulator that develops guidance, standards, and expertise in partnership with industry to strike a balance between innovation and safety.” (p. 83).

¹⁹ On the issue of the use of algorithms in the US justice system, see Kehl, Guo and Kessler, 2017. The potential implications for consumer law have not really been examined yet.

The relevance of the question whether and to what extent developments in artificial intelligence will affect the practice of scoring is probably fairly limited, because the difficulties involved in gaining insight into algorithmic decision-making processes are not confined to ‘modern’ processes. Even quite conventional algorithms based – like the Schufa credit score – on multivariate or non-linear regression models, for example, are not immediately decipherable even to specialists (Lipton, 2017). It is not only in the recent past, then, that the black box which is often cited in connection with artificial intelligence has posed a challenge. On the contrary, it is not much of an exaggeration to say that the black box has accompanied the development of software from the outset, a point which is made incisively in Passig, 2017.

The disclosure of source codes, even those of simple computer programs, is not generally of much help to consumers (see also section B.VII.2 below). Most data subjects are not computer specialists. And even if they were, the technical complexity of the computer systems whose decision-making behaviour is to be made ‘transparent’ necessitates a different form of transparency from the mere disclosure of programming codes, even for the purpose of verification by experts (Wischmeyer, 2018; Samek, Wiegand and Müller, 2017; Selbst and Powles, 2017; Montavon, Samek and Müller, 2018; Gigerenzer, Wagner and Müller, 2018; *Gesellschaft für Informatik* (German Informatics Society), 2018). This appraisal forms the basis of a research programme which is currently being very vigorously pursued and is examining how the way in which complex algorithms work can be explained to people comprehensibly. In the field of artificial intelligence, the key concepts in this discussion are interpretable machine learning and *explainable artificial intelligence* (XAI – see *Gesellschaft für Informatik*, 2018; see also Wachter, Mittelstadt and Floridi, 2017; Selbst and Powles, 2017; Selbst and Barocas, 2018).

An adequate understanding of transparency in this context is one that seeks to embed algorithmic decision-making systems in explanatory and review structures based on a division of labour (Wischmeyer, 2018). To be transparent, a system of algorithmic decision-making need **not** be visible to the observer in all its details – which, in the case of neural networks, would certainly be difficult (see, for example, Ribeiro, Singh

and Guestrin, 2016; Burrell, 2016; Alber, Lapuschkin and Seegerer, 2018) – but must be explainable to its users.

The central question is “Is it possible to explain or make visible retrospectively how the result was arrived at?” (Passig, 2017, p. 25). To achieve transparency, then, it is not necessary to establish a “full understanding” of scoring software in all its details. It is sufficient to create means of obtaining knowledge of the way in which an algorithm works. Even in conditions of incomplete transparency and incomplete comprehension of scores, testing of their functioning is possible by calculating scores for exemplary cases. This method is called *black-box tinkering* (Perel and Elkin-Koren, 2017; Wachter, Mittelstadt and Floridi, 2017). A proposal for a “transparency interface” (Gigerenzer, Wagner and Müller, 2018) follows this methodology, as do the proposals made by the German Informatics Society (*Gesellschaft für Informatik*) that algorithm testing be made into a robust regulatory instrument (*Gesellschaft für Informatik*, 2018, which also recommends the creation of a right to conduct tests). To this end, the input to scoring systems (AI in general) would be systematically varied and the output evaluated. This could be required, for example, by a supervisory data protection authority in the framework of data protection audits under Article 58(1)(b) of the GDPR and possibly be conducted by that authority itself. Although this would not necessarily make the detailed internal operations of the black box recognisable, it would provide sufficient knowledge of the **relevant** workings of the algorithm. This testing, by the way, is in line with the logic of Stiftung Warentest, the German Comparative Testing Foundation. The Foundation does not study architectural plans or recipes but draws conclusions about the relevant attributes of a product from its systematic use.

Developers and users of scoring systems have also come to appreciate the importance of the traceability of their methods. Francesca Rossi of IBM put it this way in an interview with a German daily newspaper: “Besides deep learning, there are systems like decision trees, which are easier to retrace but unfortunately not quite so accurate. So we have to find out which is more important to us: the accurate outcome or the traceability of the process.” (Rossi, 2018).²⁰

5. Transparency as a condition for a social debate on scoring

Inherent in the use of scores is the risk of lending the appearance of objectivity to judgements that are insufficiently discussed within society and so placing them beyond criticism (for a fundamental treatment, see Porter, 1995; see also, for example, Heintz, 2007). This criticism of quantification and of the “social use of numbers” (Vormbusch, 2012, p. 37) has become the subject matter of numerous studies, which constitute a productive field of research. Established reference areas for such analyses are economic policy (Weingart and Wagner, 2015; Wagner, 2018; Schlaudt, 2018) as well as various fields of education, health and social policy (Muller, 2018) and in particular the actions of international organisations where these are substantially based on indicators, rankings, thresholds, etc. (Davis and Fisher, 2012; Rotenburg, Merry, Park and Mugler, 2015; Merry, Davis and Kingsbury, 2015; Merry, 2016).

With the aid of numerous examples and case studies, these analyses have highlighted that, wherever there is a decision to be taken, it is considerably more convenient to proceed from a numerical value than from a multi-layered, sophisticated and possibly ambivalent assessment of a fact or – as in the case of scoring – a person. Scores go almost as far as is possible to reduce the complexity of judgements. This makes the use of scores in decision-making very appealing, especially if the decisions are taken, or have to be taken, in automated form, rapidly and in huge numbers. The judgements that have to be made in the development of scoring methods are often not discussed in a way that is conducive to the subsequent social use of those methods. The criteria to be included in a points system for scoring healthy lifestyles in a health insurer’s bonus programme and the weight to be given to each criterion do not usually attract public notice (see also section C.III.2 below). It is a moot question, for example, whether bonus points should be awarded solely for activities that benefit the scored person’s state of health or should also be credited for those that help to make the health system work better, such as blood donation and bone-marrow typing, or even for activities which are not health-related but which are deemed socially valuable, such as voluntary work. The absence of public discussion on the judgements that have to be made when creating a scoring system could be described as a lack of politicisation, that is to say the imposition of normative opinions with considerable social consequences without a preparatory and accompanying social debate.

²⁰ On the question whether this trade-off actually exists in the reality of scoring, see, for example Hand, 2006, who advances good arguments against its existence. The point here is only that practitioners are evidently recognising the interest in comprehensibility. Accordingly, it is not a rebuttal to state that random forests, in other words a large number of decision trees (*Gesellschaft für Informatik*, 2018), are not necessarily easier to interpret (Groll, Ley, Schauburger and Van Eetvelde, 2018) and that the comprehensibility problem will not therefore disappear or is even necessarily be reduced as a result of the change of method.

Insufficient transparency, moreover, may reinforce misconceptions within society as to what a score actually signifies (see also section C.III.3 below). A score awarded by a motor insurer as the basis for a telematics-based tariff may be structured in such a way as to cover not only driving habits that can be influenced, such as the care with which the driver brakes and accelerates or observes speed limits, but also factors that are not related to driver performance but influence the likelihood of an accident. These could, for example, be the ratio of urban to rural drives, since accidents are more likely to occur in towns and cities, or the ratio of night-time to day-time drives, because driving at night increases the probability of having an accident. If the scored driver or the public at large have the impression that the score is primarily an indicator of driving skills, a gap opens up between the real significance of the score figure and its social use. The purpose of transparency in this context is therefore to ensure that the meaning of scores is realistically appraised and that scores are used only to convey that meaning.

II. Non-discrimination and equal treatment

Scoring processes result in different scores from one person to another. That, indeed, is their very purpose, for scores mark differences, and the aim of scoring systems is to differentiate. At the heart of every economic and legal order based on market economics and democracy is scope for private autonomous differentiations. In principle, every merchant is free to decide whether to conclude a contract with a consumer. A contracting obligation exists only in a few exceptional cases. Conversely, protection against discrimination is continually gaining ground within the legal order. Scoring systems operate in precisely this field of tension between entrepreneurial freedom and social values, the balance between which requires constant readjustment.

1. What is discrimination?

The phenomenon of discrimination is understood in a broad sense in this report. It encompasses actions and structures that lead to those who possess particular characteristics²¹ – such as women, homosexuals or people of ‘alien’ ethnic origins – being disadvantaged within society. A guide to the characteristics that are relevant in this context is given in section 1 of the General Equal Treatment Act (*Allgemeines Gleichbehandlungsgesetz*), which states that “The purpose of this Act is to prevent or to stop discrimination on the grounds of race or ethnic origin, gender, religion or belief, disability, age or sexual orientation”. In other instruments the set of ‘discrimination grounds’²² is defined differently, although there are, of course, numerous overlaps. In each case, the key characteristics on which discrimination should not or must not be based are the result of social nego-

tiating processes, of insight into mechanisms of social exclusion and into historical injustice and, ultimately, of civilisatory progress (Fritzsche, 2017). They need not be defined identically for all areas of life and social situations and they are open to legislative adaptation and development.

Such a definition of the phenomenon of discrimination is to be distinguished from two competing meanings of the term: on the one hand, not every distinction made between individuals *per se* is discrimination within the meaning of this report. Such an interpretation of the concept (cf. Adomeit, 2002, and Picker, 2008) would imply the inclusion under the heading of discrimination of numerous social interactions which are not an issue and which do not create any need for political action, extending even to a restaurateur’s differentiation between customers who are willing to pay and those who are not. Some specialised statistical terms, such as discriminant function analysis, are based on a value-free understanding of the verb to discriminate. On the other hand, our use of the term ‘discrimination’ is not meant to imply that distinctions made on the basis of the discrimination grounds are socially unacceptable, let alone legally prohibited.

On the contrary, the term ‘discrimination’ is intended to designate any unequal treatment based on a criterion that is held to require special legitimisation if used as a ground for differentiation. A mere reference to the free choice of the person making the distinction does not suffice to justify the unequal treatment.²³

21 For the avoidance of doubt, these characteristics are not objective attributes that are inherent, as it were, to the person subjected to discrimination. Non-discrimination rules afford protection against distinctions made on the basis of purely attributed characteristics (Schiek, 2000; for the General Equal Treatment Act (*Allgemeines Gleichbehandlungsgesetz*), see Bundestag printed paper 16/1780, pp. 30–31). This explains, for example, the prohibition of discrimination based on ‘race’, which is a social but not an anthropological category.

22 A wide and potentially confusing terminological diversity prevails in this field. Reference is made not only to *Diskriminierungsmerkmale* (“discrimination grounds” – Pärli, 2017, pp. 106ff.) but also to *verbotene Merkmale* (“prohibited grounds” – Schramm, 2013, p. 7), which does not mean, of course, that the grounds themselves are prohibited but rather discriminatory treatment based on those grounds in certain circumstances. The same meaning is given to the term *geschützte Merkmale* (“protected grounds” – Schramm, 2013, p. 3 *et passim*); this term expresses that the purpose of the discrimination ban is to protect those who possess particular characteristics from discrimination on those grounds.

23 From a legal perspective, the unequal treatment on the part of the decision-maker is not conclusively legitimised by the reference to personal autonomy, understood as “the principle of the individual shaping legal conditions according to his or her will” (Flume, 1965, p. 1) and as “recognition of the autocracy of the individual” (*ibid.*, p. 6).

2. Discrimination through scoring input

The risk of discrimination is inherent in scoring systems. Scoring processes use a number of a person's attributes in order to obtain a score for that person. If these attributes that are recorded and used to calculate a score include membership of a protected group, it goes without saying that the scoring process will have discriminatory effects. We can call this direct discrimination through scoring. For example, a scoring system discriminates directly if gender or membership of a particular ethnic group is part of the scoring input. The existence of such a direct discriminatory effect of scoring is comparatively easy to ascertain. If it can be ascertained (see chapters I.1 and I.2 above on the transparency of scoring methods) that an attribute which is a discrimination ground is part of the input for the calculation of a score and that this specific attribute worsens the person's score, discrimination is duly demonstrated (see also Hacker, 2018).

Discrimination risks in scoring, however, may result from something other than a scoring parameter being derived directly from a protected ground. That, indeed, is most likely a comparatively rare phenomenon. The more socially relevant risk is that of indirect discrimination by scoring systems, which is informatively depicted by Hofstetter, 2016. In such cases, the scoring method involves input criteria which, though innocuous in themselves, are statistically connected with protected grounds. For example, body size, consumption patterns or customary leisure activities can be used indirectly to obtain an entry under the 'gender' heading. Illustrative material on this phenomenon is to be found in those studies that reveal how characteristics like ethnic origin or sexual orientation can be obtained with some degree of probability from the narrow information base of a few clusters of Likes on Facebook (Kosinski, Stillwell, and Graepel, 2013; see also section B.

VIII.2 below). It is therefore plain to see that direct discrimination does not account for the whole problem of discriminatory scoring systems.

By now we have reached the point where the discrimination issue becomes extremely complex, for there will be scarcely any attributes which may safely be assumed from the outset not to be linked with the existence of a discrimination ground. On the contrary, numerous attributes that go to make up a score correlate with discrimination grounds, sometimes positively and sometimes negatively, sometimes more strongly and sometimes less so. It is not unreasonable to assume, for example, that men are more likely to take risks when driving, which leads in turn to a comparatively poor score from motor insurers. Accordingly, the link with the neutral criterion of driving behaviour at least merits attention in the light of protection against discrimination (for more on this example, see item 3.2 below).

In actual fact, the issue is even more complex, because discrimination problems also arise if there is a link with an attribute which, though not connected to any statistically significant extent with any individual discrimination ground, such as gender or ethnic origin, is linked to combinations of various discrimination grounds, as in the case of persons of a particular sex who also have a particular ethnic origin. As a result of intersectionality research, an academic discipline has been established which deals with the situation of persons who belong to two or more protected groups (Meyer, 2017, and Chege, 2012). Studies in this field are designed to demonstrate, with the aid of numerous examples, that such combinations of protected grounds form separately contoured discrimination categories which cannot be reduced to the 'simpler' discrimination grounds.²⁴

²⁴ Anti-discrimination law as it stands does not yet make provision for this situation. The Court of Justice of the European Union, the foremost instance for the interpretation of anti-discrimination law, which is harmonised throughout the EU to a considerable extent, does not recognise intersectional discrimination as a separate category of offence; see ECJ judgment of 24 November 2016, *Parris*, C-443/15, EU:C:2016:897, paragraphs 80 to 82). In view of the criticism and lively discussion among scholars and practitioners about the classification of intersectional discrimination, it seems likely to be the subject of more proceedings before the ECJ.

3. Score quality and non discrimination

An analysis of discriminatory scoring that focuses on input into the scoring process concerns both behaviour-modifying and behaviour-predicting scoring systems. Only in the case of predictive scoring is the quality of the score in the spotlight. Such scoring systems are designed only to perform a forecasting function. They may be better or worse at performing that function. There is no evident relationship between the high quality of a score and the prohibition of discrimination. As requirements for a fair scoring process, they may go hand in hand, but they may also be at odds with each other.

3.1 Score quality and non-discrimination in harmony

The regulatory aims of quality assurance and of a general prohibition of discrimination go hand in hand in a particularly intelligible way in cases where the link with a protected ground is merely a reflection of conventional preferences and aversions. Such “preference-driven discriminations”, as they are described (Gardner, 1998; Block, 2018, on section 3 of the General Equal Treatment Act, points 10ff.), contribute nothing to the good predictive performance of a scoring method. On the contrary, in such cases, the scoring user’s prejudices distort its appreciation of the scope for a more effective scoring method. To take a hypothetical example, if the deviser of a scoring system believes that “women are basically worse drivers” and therefore systematically deducts points on the basis of a driver’s sex, he is not only discriminating but is actually designing his scoring method less well than he could, for without any empirical basis the scoring method imputes a greater probability of accidents to women drivers than to their male counterparts. Such a scoring system is flawed. This caricature should suffice to show that today’s urgent practical problems relating to discrimination lie in another direction, because far more relevance attaches to the “discriminating flaws” in scoring systems. By this we mean that the quality of a scoring system is not equally high or low for all scored persons but that defects in scoring systems have a particularly significant impact on certain groups of people. Let us just imagine

a toaster that goes on fire more often when operated by women than when operated by men – this has always been an absurd notion, but in the world of complex predictive scoring algorithms it becomes a real problem. As an individually focused ‘product’, scoring systems will work with varying degrees of efficiency for different people. There are some parallels to be seen in the discussion as to whether particular medicines demonstrate avoidable quality variations when taken by men and women because of the social conditions in which pharmaceutical companies research, develop and test them – for example if only male test subjects have been used (on this point, see Nieber, 2014).

The discussion on ‘big-data discrimination’ is generating important insight into the interaction of quality assurance and non-discrimination (Barocas and Selbst, 2016). The discussion has highlighted how the flawed construction of big-data applications can produce discriminatory effects. A recent summary (Executive Office of the President of the United States, 2016) divides the causes of discriminatory effects into those inherent in the data used as inputs into big-data applications and those related to the inner workings of the algorithm with which these inputs are processed (see also Hacker, 2018).

One example would be an automated system for the selection of job interviewees which always awarded the highest aptitude ratings to male applicants from non-migrant backgrounds because the system had ‘learned’ its decision-making criteria from the company’s historical data, which marked out men from non-migrant backgrounds as particularly successful (an early, much-discussed scenario presented in Lowry and Macpherson, 1988). The “unintentional perpetuation and promotion of historical biases” (Executive Office of the President of the United States, 2016, p. 8) produced by this body of data is the most likely explanation in this case.

3.2 Score quality in conflict with non-discrimination

Ensuring a high quality of scoring and providing protection against discrimination do not necessarily or invariably go hand in hand. A particular attribute may have predictive power but at the same time it may itself be a protected ground or correlate with the existence of a protected ground. The example of the risky driving which is allegedly most prevalent in men has already been cited. At the same time it is evident that the inclusion of risky driving will considerably improve the predictive power of a score that is designed to forecast the probability of an accident. The predictive performance of the scoring method and non-discrimination may therefore become conflicting aims.

This dilemma has to be recognised. The problem of poor score quality is not to be equated with that of discriminatory scoring methods. The question of striking the right balance must be debated within society. It could be a heated discussion, but it need not be. In the case of drivers being marked down for risky driving, the question probably answers itself: surely no one would take the view that risky driving should be omitted from the scoring inputs just because men, as the riskier drivers, would be hit harder than women, who drive more carefully. Behind that answer lies the legitimate social expectation that men should be capable of changing the way they drive. For other grounds, however – such as sex or ethnic origin – the same expectation would have to be dismissed as unreasonable to unfulfillable.

It would be too simplistic by far to argue that discriminatory effects should be accepted without demur for the sake of the predictive performance of a scoring method. To do so would be to disregard the fact that one of the factors behind predictive effectiveness may be the result of previous group-based discrimination (Britz, 2008). This very state of affairs would be perpetuated by the recognition of directly discriminatory predictive scoring systems (Block, 2018, on section 3 of the General Equal Treatment Act, point 15). This connection may be illustrated by means of an example: it seems possible that membership of a particular ethnic group possesses predictive power when it comes to determining the probability of default on a loan. It would, however, be wrong to dismiss out of hand the suspicion that this higher probability reflects the fact that the economic circum-

stances of members are generally more straitened because of past discrimination. If it is accepted that ethnic origin is a valid scoring category for decisions whether to grant loans, such conditions of social disadvantage will be further entrenched (Kim, 2017).

3.3 The example of ‘redlining’

The use of address details for the purpose of credit scoring offers an insight into the conflict between the predictive quality and the anti-discrimination credentials of scoring models (see also section B.V.2 below). Section 31(1)(3) of the Federal Data Protection Act stipulates that predictive scoring is admissible only if “other data in addition to address data are used to calculate the probability value”. This provision outlaws the practice known as ‘redlining’. “The term denotes the practice of circling parts of a map with a red line to define areas to whose residents, for example, a bank will not issue any mortgages or will only issue them on less favourable conditions. In redlining, then, an appraisal of the data subject’s address may suffice for a loan to be refused or to be granted on less favourable conditions or for the imposition of more stringent terms of payment, such as cash before delivery” (Hammersen and Eisenried, 2014, p. 343).

One argument which is sometimes advanced for the prohibition of redlining is simply that the predictive power of address data is questionable: “As long as it is not known who lives in the penthouse and who in the basement, predictions will remain fuzzy” (von Levinski, 2018, on section 31 of the Federal Data Protection Act, point 42.1). To the extent that this is true, the provision prohibiting the sole use of address data does not raise any particular difficulties, because the legislator has ‘only’ banned scoring agencies from basing their scoring methods decisively on a data category which, when all is said and done, can contribute nothing to a good predictive output. The point is, of course, that we cannot rule out in advance the possibility that a person’s address might be an input which permits an adequate prediction of the probability of default on a loan. The prohibition of redlining thus has a genuine fairness dimension, and quality assurance is not its sole purpose.

The aim of the prohibition is to halt the momentum of an assumed self-fulfilling prophecy, namely “Bad risk ratings lead to worse terms and conditions and higher financial burdens, which tend in turn to thwart the fulfilment of payment obligations” (Korczak and Wilken, 2008, p. 24; see also Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014).

Redlining could well have a trend-setting effect. Every use of an attribute for the purpose of predictive scoring may be liable to set a process in motion for those who possess that attribute, leading to the entrenchment of social disadvantage, for which reason it must be prevented (Kim, 2017).

3.4 Good scoring in conflict with non-discrimination

It is not enough for a scoring model to display a high quality level. That quality must also benefit all scored individuals and groups in equal measure; it should be ‘non-discriminatorily good’. This requirement brings us to the lively discussion conducted predominantly by IT specialists under the heading of ‘algorithmic fairness’ (*Gesellschaft für Informatik*, 2018). In the debate on algorithmic fairness, the fairness requirements that may be made of predictive scoring models, among other things, are formulated mathematically. A large number of criteria – ‘measures of fairness’ – have been developed in the course of this discussion, and these can be used to assess the extent to which a scoring model may be described as fair (for more details, see chapter B.VI below). This discussion analyses where measures of fairness are mutually contradictory and cannot be simultaneously achieved (for an introduction to the debate, see Zweig and Krafft, 2018).²⁵ Which measures of fairness are more important than others and must therefore be adhered to by developers of predictive scoring models is a question that cannot be answered on the basis of mathematical criteria. Decisions on fairness priorities, which may be ‘tragic choices’, the term used by Calabresi & Bobbitt (1978), must be discussed by society, and the legislator, if necessary, must enshrine them in binding provisions.

²⁵ This dilemma is illustrated by the COMPAS algorithm which is used in the US criminal justice system and which has been much discussed in Germany as in other countries. COMPAS predicts how likely it is that a defendant will re-offend within the next two years (Moll, 2016). To this end, 137 questions about a person are answered with information collected either at a direct interview or from police records; these answers are compared with databases on convicted offenders, and a risk score is calculated. In 2016 the New York-based non-profit newsgroup ProPublica demonstrated in an extensive study that the COMPAS algorithm discriminated against Afro-Americans in that it falsely flagged them as likely re-offenders far more often than it wrongly predicted the actual re-offending rates of white people, such overestimation errors being classed as ‘false positives’ (Angwin, Larson, Mattu and Kirchner, 2016). Some researchers have responded to the discrimination charge by stating that there is an equal probability of re-offending for defendants with the same risk score, regardless of whether they are Afro-American or white (Angwin and Larson, 2016).

4. Undesirable unequal scoring-based treatment beyond discrimination

Discrimination must be distinguished from other forms of undesirable unequal treatment. This means forms of differentiation which consumers perceive to be unfair or inappropriate and which therefore, in their view, ought not to be used as decision-making criteria. Normally, the law as it stands does not encounter problems in this respect; the body of private law is interpreted in the vast majority of cases to the effect that decisions affecting private individuals need not fulfil heteronomously prescribed rationality criteria other than in exceptional cases. Although it should not be forgotten that private law is permeated by numerous equal-treatment requirements in particular areas – from employment through capital markets to competition (for a detailed examination, see Grünberger, 2013), being treated differently from someone else “without reasonable grounds” cannot be regarded as a problem in private law in normal circumstances, because in private law the will of the participating parties is considered paramount rather than ‘objective’ rationality criteria. Private autonomy means precisely the freedom of a decision-maker not to be accountable to the law, however much others might regard that person’s decision-making as idiosyncratic, unwise or morally repugnant. To quote the Federal Constitutional Court, there is no “objective constitutional principle according to which legal relationships between private actors would be generally subject to equality guarantees. In principle, all persons have the freedom to choose – according to their own preferences – when, with whom and under what circumstances they want to enter into contracts”. (Federal Constitutional Court, Order of the First Senate dated 11 April 2018 in case No

BvR 3080/09, headnote 1). It is a moot point whether this perspective requires adaptation and further development in the light of ubiquitous data collection and personality profiling, because it may come into conflict with consumers’ expectations in terms of fairness. From a consumer’s point of view, the brand of mobile phone he or she uses to book a hotel room should not affect the price of the room in any way whatsoever (Townley, Morrison and Yeung, 2017; Zander-Hayat, Reisch and Steffen, 2016; Zander-Hayat, Domurath and Gross, 2016). The posts on social networks to which a consumer gives a Like or which he or she shares or comments on should not receive any consideration when that consumer’s creditworthiness is assessed. In relation to scoring, this means that if input used to calculate a score is perceived to be inappropriate by the scored person or by the general public, it should at least raise the question whether the regulatory framework should tolerate this state of affairs or whether it has to be remedied.

The issue here, however, is not that of discrimination as presented at the start of this part, because the distinctions that are made in this case are not based on the ‘classical’ discrimination grounds. The users of Apple products, for instance, could scarcely be described as a typically disadvantaged group. On the contrary, the question to address will be where, in the digital world in general and in the realm of scoring in particular, do we find macrosocially legitimised limits of unequal treatment which are not covered by existing anti-discrimination law.

III. Enforcement of rights

Fairness requirements relating to scoring that are established in law cannot be fulfilled unless there are effective official control structures in addition to instruments for the enforcement of rights in civil law. In its report entitled *Consumer Rights 2.0*, the SVRV explained in detail that it is not feasible to place responsibility for enforcing rights, particularly those relating to fair competition, standard business terms and non-discrimination, entirely in the hands of consumers (SVRV, 2016). Several reasons can be cited for this finding (see also Podszun, Busch and Henning-Bodewig, 2018). In judicial proceedings consumers are typically in a position of *de facto* inferiority in relation to a corporate party. Reasons for this include consumers' more restricted budgets and more limited knowledge of the law but also their relative inexperience of the judicial process (Fries, 2016). In addition, the monetary amounts at stake in consumer cases are usually low. For this reason consumers often have only a minimal economic interest in the subject of the dispute and hesitate to incur the trouble and risks associated with the enforcement of individual rights. It fits perfectly into this picture that spectacular litigation successes on the part of individual consumers – one need only think of the Schrems case for enforcement of the guarantees provided by EU data protection legislation (ECJ judgment (GC) of 6 October 2015, C-362/14, EU:C:2015:650; for more on this judgment, see SVRV, 2016, pp. 60 and 70) – are very clearly attributable to actions brought in the public interest, not to the cost-benefit considerations of one consumer focused on his or her individual gain.

Even in systemic terms, an individually centred regime for the enforcement of rights (Hellgardt, 2016, pp. 549ff. and pp. 560ff.) seems unsatisfactory (SVRV, 2016). The considerable length of time that elapses between the first occurrence of a problem in the everyday lives of consumers and its judicial resolution, the cross-border nature of numerous legal conflicts in the digital world (for a detailed treatment, see Calliess, 2006) and the de-territorialisation of the law on the worldwide Web are arguments against leaving the enforcement of rights primarily to the aggrieved consumer himself. A consumer, moreover, will only recognise and judge a narrow, individualised segment of corporate activity. For example, advertising, contract offers, prices and contractual terms and conditions can be tailored to individuals by means of algorithms. Whether unlawful discrimination

underlies this individualisation can scarcely be reliably established as long as the situation is considered from the perspective of the individual consumer (see also section E.III.5 below).

Private consumer associations do not share the aforementioned weaknesses of an individual consumer. Their mere institutional status, however, does not enable them to enforce rights. They lack the powers of state authorities, which can require companies to disclose information, which possess extensive rights of investigation and intervention and which, in the European Union, are part of international networks of public authorities and are therefore more easily able to exchange information and share experience (SVRV, 2016).

Enforcement of legal requirements for fair scoring is therefore a task that must not be imposed on the consumer alone but must also be entrusted to governmental bodies. This observation turns the spotlight on questions relating to the responsibilities, organisation and staffing of the governmental bodies that might fit the bill. In connection with the enforcement of numerous consumer rights in the digital world, the SVRV (SVRV, 2016) presented models from other countries – the United States, Britain and the Netherlands – and outlined the specialised skills and expertise that a German institution would have to establish and nurture to enable it to enforce these rights effectively. The SVRV believes that there is an urgent need for action to create such a regulation agency for digital operations (SVRV, 2016; see also BMWi [Federal Ministry for Economic Affairs and Energy] and BMJV [Federal Ministry of Justice and Consumer Protection], 2015; Bundestag printed paper 19/1982, p. 8) and has presented options for such a course of action.

IV. Score quality

1. Quality of the algorithm underlying a score

There is a divergence between the statutory requirements for a scoring system, which are based on the largely undefined legal terms in section 31(1)(2) of the Federal Data Protection Act (for more details see section E.III.3 below), and the standards discussed by academic scholars in the field of empirical social research. From a legal point of view it is remarkable that the courts have had little opportunity yet to flesh out the undefined terms. Consumers are evidently not or only very rarely prepared to enforce their claims through the courts. Without pertinent judgments, however, the arguments needed to underpin the openly worded statutory provisions are lacking. Regardless of the statistical models with which scoring agencies operate in practice and the quality rules they have laid down for themselves, it must be noted that the aforementioned definition of terms alone does not guarantee the fairness and quality of scoring,²⁶ not least because of the lack of judicial case law that would help to clarify precisely what is meant by scientifically recognised standards.

Here are some examples of ‘scientifically recognised’ standards that are recommended for the publication of empirical findings in the context of a peer-review procedure in the field of empirical social science: identification or transmission of all utilised input variables or a summary of them (averages, standard deviations, etc.) to enhance the verifiability of estimated findings, transparent calculation and presentation of findings (including the consumer attributes (predictor variables) and the target variable, the significance of the various predictor variables, the influence of predictor variables on the target variable and the quality of the estimation model) and assessment by qualified external consultants of findings and, ideally, of the whole estimation procedure.

As regards the calculation of scores, the following machine-learning processes are available: regression, clustering methods (e.g. k-means), decision trees, ensemble methods (e.g. boosting and random forest) and, at the test stage, deep learning in architectures such as neural networks (Lessmann, Baesens, Seow and Thomas, 2015; Thomas, Crook and Edelman, 2017).

In the realm of credit scoring, logistic regression is known to be the customary model, and newer methods such as neural networks, though tested, have not yet been used operationally (Schröder *et al.*, 2014; Thomas *et al.*, 2017; Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014).

26 Accessed at on 16 May 2018.

Background: excerpt from a highly simplified algorithm²⁷

An algorithm is not a magic formula but merely describes a process of mechanical data analysis, based on established – and often well-researched, as in the case of logistic regression – analytical procedures from the realm of statistics (James, Witten, Hastie and Tibshirani, 2013; Jentzsch, 2018). Even highly non-linear algorithms, which include neural networks, can be examined structurally and can be broadly represented increasingly well as decision trees or regression equations (cf. for example Montavon, Samek and Müller, 2018). Accordingly, it is pointless to argue over the question whether neural networks can always, or only in special cases, be comprehensibly interpreted (cf. section B.I.4 above).

Logistic regression analysis tests whether a link exists between two or more independent variables (in our case consumer attributes) and a binary dependent variable (0 or 1 – in our case the target attribute). To be more precise, it tests whether the consumer attributes influence the probability that the target variable will have the value 1 or how strongly the individual manifestations of the consumer attributes influence the target variable, the strength of this influence being expressed by the regression coefficient.

Logistic regression analysis can be used, for example, to examine the following question:

“What influences the probability that a consumer will not repay his loan? (where **Y** is the risk of default)

the number of current accounts held by the consumer (**X**₁),
the number of his current loans (**X**₂), or
the consumer’s address (**X**₃)?”

Technically, this may be expressed in the following formula:

$$P(Y = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}$$

The following values can be derived from the use of logistic regression:

- the significance of the influence of a consumer attribute on the target variable (calculated, for example, by means of the Wald test); this serves as the basis for the decision whether to include a consumer attribute in the score calculation at all;
- the extent of the influence of a consumer attribute on the target variable, expressed by the regression coefficient; this is generally the basis for the weighting of consumer attributes in the score calculation;
- the quality of the overall model, also described as its discriminatory power, which may be expressed, for example, by the Gini coefficient.

²⁷ See, for example, Auer and Rottmann (2015) and Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe (2014).

As a matter of principle, it should be noted that the almost impenetrable domain of artificial intelligence does not yet play a part in scoring, at least not in Germany. This is probably due in part to the fact that the potential of artificial intelligence beyond image and speech recognition is far less impressive than the developers of AI assert. There is extensive literature showing that good scoring depends not so much on the scoring method as on the quality – or lack of quality – of the input data, including the entity recognition, that is to say the identification of persons (for an anecdotal account, see Seibt, 2018, and chapter V below), and the stability of the scored process (cf. Hand, 2005; Hand, 2006; Verbeke, Dejaeger, Martens, Hur and Baesens, 2012; and – in spite of a certain overoptimism regarding the capability of AI – Lessmann *et al.*, 2015). Verbeke *et al.* (2012) therefore reach the conclusion that easily comprehensible decision trees could be given precedence over complex methods, not least because certain decision trees achieve comparable predictive power (see, for example, Phillips, Neth, Woike and Gaissmaier, 2017) while typically requiring fewer input variables (Jenny, Pachur, Williams, Becker and Margraf, 2013; for an analytical explanation, see Brighton and Gigerenzer, 2015). According to this body of literature, besides investments in new estimation models, even more should be invested in the measurement quality of models that are already in use.

The use of comparatively simple and comprehensible scoring algorithms in practice, moreover, does not necessarily mean that the utilised algorithms would be rendered transparent – which, from a scientific perspective, could easily be possible.

Which algorithms are used in telematics-based motor insurance is not evident from publicly accessible sources, an issue which is addressed in Part C below. In the calculation of health scores for health insurance policies, scarcely any use is made of complex algorithmic decision-making methods, the calculation of total bonus points in the statutory health insurance system being based on a simple addition of individual bonus points. Other providers in the secondary healthcare market such as *Vitality*²⁸ and *Dacadoo*²⁹ have more sophisticated scoring systems.

The described models serve, with the aid of numerous consumer attributes – such as place of residence, current loan agreements, etc., in the case of credit scoring, or braking and acceleration behaviour, driving times and places, etc., in the case of telematics-based motor insurance – to predict consumer behaviour and/or the potential consequences of consumer behaviour such as the probability of defaulting on loan repayments and/or the probability of being involved in a road accident. These potential outcomes are known as target variables.

How ‘good’ an algorithm may be in practice depends on the selection of the algorithm itself, the utilised target variable, the selection of the consumer attributes and other model parameters, such as the relative weighting of overoptimistic and overpessimistic predictions. A model may be described as very good if, in the vast majority of cases, it puts consumers into the correct risk category, in other words classifies them correctly as ‘good’ or ‘bad’ risks in terms of the probability of defaulting on a loan or having an accident.³⁰

In the context of score calculations, three categories of quality measurement are available: measures such as the Gini coefficient or the area under the curve (AUC), which describe the discriminatory power of a model,³¹ measures such as the Brier score, which indicate the accuracy of the predictive power of a model, and classification error (see, for example, Lessmann *et al.*, 2015).

28 https://www.generalivitality.de/vmp/punkte_und_status/tipps_zum_punktesammeln; accessed on 24 May 2018.

29 <https://www.dacadoo.com/?lang=de>; accessed on 24 May 2018.

30 It should be said by way of qualification that there are no completely infallible methods and decisions. This applies to algorithm-based decisions and to human classifications, such as those taken by credit clerks in banks.

31 A perfect prediction instrument would have the Gini coefficient 1.00. A coefficient of 0.00 means that the predictions are pure guesswork.

From a consumer-policy perspective, the problem is essentially that score providers' quality measurements are not reported, nor are there statutory minimum requirements for the quality of scoring methods. The minimum requirement in law, set out in section 31(1)(2) of the Federal Data Protection Act, is already satisfied by the mere use of one of the aforementioned algorithmic decision-making procedures – as weak a quality requirement as any empirical social researcher could imagine (for more details, see section E.III.3 below).

Another possible problem arising from a particular choice of consumer attributes for the calculation of scores is a phenomenon known as the multicollinearity of data. Depending on which attributes are used to calculate a score, a close correlation between individual attributes cannot be ruled out. In the case of a credit score, this might apply, for example, to the number of current accounts held by a consumer and his level of indebtedness. In a hypothetical scenario, consumers with many current accounts would tend to have more debts. If these predictor variables correlate with each other, it can complicate the task of determining the actual contribution of one of them to the statistical prediction of the target variable. In other words, because it is statistically difficult to distinguish between predictor variables, the algorithm attributes changes in scores to changes in debt levels, even though they actually result from changes in the number of current accounts, or vice versa (see also Auer and Rottmann, 2015). In theory, multicollinearity could even reverse the direction of the influence exerted by a predictor variable on the score within the statistical model; for example, although an attribute might have a benign influence on a person's creditworthiness, collinearity could cause the algorithm to assess a detrimental effect (Schröder *et al.*, 2014).

Another aspect relating to the quality of an algorithm is possible deterioration in its predictive power as a result of structural changes, described as exogenous effects. It may be assumed, for instance, that a series of social changes occurring at lengthening intervals will diminish the stability of the statistical correlations established in scoring models (Hand and Henley, 1997). The predictive power of a statistically calculated score may therefore change in the course of time. Recurring cyclical fluctuations can also influence the reliability of credit scores (Schröder *et al.*, 2014). In the healthcare sector a structural deterioration in health scores could occur as a result of demographic change and the consequent decline in the average state of health of the population. Scoring models must therefore be safeguarded at regular intervals against the influence of exogenous effects. In addition, social norms inherited from old data, such as the prevalence of male chief medical officers, may be encoded and involuntarily perpetuated through models based on data that have not been updated (see, for example, Lowry and Macpherson, 1988).

2. The utility of newer and more complex algorithms

Score providers are free to make changes to their existing algorithmic decision-making procedures at any time, for example by replacing the data categories they use or adapting their weighting parameters, or to switch to a new decision-making procedure altogether, for example by changing from logistic regression to a new machine-learning process (see also section B.I.4 above).

In the motor insurance industry, for instance, an insurer reserves the right to alter its weighting in the light of new findings in the field of accident research or the results of a completed pilot project,³² which means changing the scoring rules, for example by adapting the range of input data or adjusting their relative weighting, and consumers may have to adapt their behaviour to a new assessment template, provided they are even aware of the change. Algorithm adaptations are welcome in principle, especially if they contribute to better score quality. A recent analysis of the driving behaviour of participants in HUK Coburg's Smart Driver programme revealed that a "clear correlation between greatly exceeding the speed limit and frequency of accidents" emerged only in cases where the statutory speed limit was being exceeded by 30 kph or more. A lower margin of excess speed had previously been imputed.³³

With regard to the switch to novel algorithmic decision-making tools such as neural networks, the criticism is made that these are something of a 'black box', even for their own developers, and are more difficult to track than comparatively straightforward estimation methods such as logistic regression, as analysis of the code serves no useful purpose (16. TB Hess LReg. 2003, 21; 17. TB Hess LReg. 2005, 10. (16th and 17th activity reports of the *Land* Government of Hesse); cf. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). There are, however, new processes for shedding light on the black box by means of systematic testing (Gigerenzer, Wagner and Müller, 2018).

New methods are not an issue in themselves; the problem lies in the capabilities of the supervisory authorities, such as the Data Protection Commissioners of the *Länder* and the Federal Insurance Office, which – unlike BaFin, the Federal Financial Supervisory Authority – scarcely possess the technical and human resources they need to be able to audit complex algorithmic decision-making processes.

³² <https://www.sparkassen-direkt.de/telematik/faq/>; accessed on 15 May 2018.

³³ Written statement from HUK Coburg dated 10 September 2018 (on the SVRV files).

V. Baseline data

1. Accuracy, currency and completeness

Errors in the baseline data may directly affect the functional quality of an algorithm, both for the computation and for the operational application of a score. In the most extreme case, the score assigned to a consumer is simply wrong because one person has been confused with another. We shall deal in more detail in the following sections with quality problems relating to input data used in the calculation of a score. In the first instance we shall examine the sources of error in the application of a score, which has not been the subject of much research.

Possible sources of error in the application of credit scores which can be gleaned from specialised literature and from consumer complaints include missing data, outdated information and mistaken identity (see, for example, Schröder *et al.*, 2014). In the realm of telematics-based motor insurance, it is reported that data have either been wrongly recorded or that errors have been introduced through linkage with flawed data. Consumers have been scored too low, for example, because of outdated maps (see Part C below). In these cases, speeding was wrongly recorded when the driver was adhering to the speed limit in an area where, some time ago, there had been road works accompanied by speed restrictions.

With regard to the quality of entity recognition, i.e. the assignment of the right data to a person, which repeatedly features prominently in anecdotal evidence reported in the press (cf. for example Seibt, 2018), there is little research material. Verbeke *et al.* (2012) highlight the importance of data quality in general

from algorithm developers' point of view and attach higher priority to better data quality than to more refined scoring algorithms³⁴ Britz (2008) reports on the significance of pure and simple input errors, though without citing individual evidence.

For reasons of data protection, as little personal information as possible is collected. This creates challenges when it comes to matching a person unequivocally with a set of data. The challenges are even greater in the case of people who have come to Germany from countries such as Bulgaria, Russia, India or Thailand, whose names on their personal documents must be transcribed into Latin script. And it seems logical, for instance, that people moving home to a different locality might cause numerous problems of entity recognition in cases where a score provider has to start by establishing a new data set, which, because of a possible lack of data linkage, contains no information about a person's credit history. If that person's creditworthiness then has to be estimated by means of geo-scoring, it may be that he will be assigned an unduly poor score (see also section B.V.2 below).

Credit scorers are particularly familiar with the challenges, which are addressed on a daily basis, as in the case of the General Credit Protection Agency (Schufa), by a large team of specialists. The number of queries and complaints received by Schufa, moreover, is extremely small – and the subject of queries and complaints is certainly not confined to mistakes in entity recognition. Accordingly, entity recognition problems cannot be all too great in number, but every time they occur they feed the bad reputation that scoring has among many people.³⁵

34 In the case of machine-learning technology, errors in the training data are also particularly irritating, because by creating an inaccurate model they distort all predictions, even if the various items of input data fed into the model are correct.

35 A further impression is obtained if one considers, for example, the number of cases handled by the Schufa customer service centre for private individuals. The Ombudsman's activity report contains the following account: "About 1,000 new questions, comments and even complaints from consumers are received daily about information stored in the Schufa database. In the [...] department, a great deal of the focus is on the accuracy of data. Four teams with a total of 75 desk officers, who include law graduates, legal secretaries, notarial secretaries and bank clerks, check whether the customers' information and comments are correct and whether entries may have to be amended or deleted" (Schufa Holding AG, 2018, pp. 36–7). If the 1,000 consumer problems per day are compared with the daily total of 400,000 or so requests made by Schufa clients (*ibid.*), a problem ratio of 0.25% emerges. This is very low, but also no doubt represents the minimum extent of the actual data problems, since in many cases these problems will probably not be noticed by scored persons.

Practitioners have made us aware of potential entity recognition problems along the whole chain of participants when Schufa scores are used. Even if the provider delivers a score that is assigned to the right person, a misassignment on the part of the requester may still occur. It is possible, for instance, that the delivered score is not correctly assigned in the decision-making process itself, since in banking and insurance, but also in the realms of mail order and e-commerce, numerous data sources are consulted when business decisions are taken. As soon as more than one data source is used, correct entity recognition can no longer be taken for granted and is open to error in the absence of an unequivocal identification number. There is plainly a need for research here so that the actual scale of the problem can be better assessed in the first instance.

2. Use of proxy variables

Paradoxically, most credit scorers, Schufa being an exception, do not possess detailed information about consumers' individual credit situations. Instead, these scorers often resort to socio-demographic and/or microgeographic data, for example to draw conclusions about the payment discipline of individual consumers from the characteristics of their residential environment (Kamp and Weichert, 2005). The payment discipline of individual consumers is deduced, for instance, from the average number of negative attributes, such as instituted debt-collection procedures, judicial debt-recovery proceedings and enforcement proceedings, in the same block of flats or the average number of negative attributes per household in the same street; this is known as geo-scoring. Although the sole use of address data is prohibited under section 31(1)(3) of the Federal Data Protection Act (see subsection E.I.1 below for more details), in proceedings against the Hamburg-based credit reference agency Bürgel in 2017, the court found that the agency had derived a customer's score from his address

alone.³⁶ There have also been reports of cases in which a consumer's forename was used to deduce his likely age (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014).

A similar case arises in telematics-based motor insurance, where night-time and urban driving often impact adversely on a person's score.³⁷ Even if it were possible to prove for the entire insured population that a causal relationship existed between the time of day and/or the location of car journeys and accident probability and even though such a relationship seems intuitive (longer reaction times and possible drink driving during the night, and greater volumes of traffic in towns and cities plus higher stress levels than on country roads), it cannot be concluded that the causal relationship increases the probability that an individual will have an accident. Particularly experienced city drivers and night-shift workers are possibly being wrongly marked down and may even be subject to direct discrimination in some circumstances (see also chapter B.II above).

Such variables that do not directly measure a consumer attribute but approximate it on the basis of other available data are labelled proxy variables. In empirical social research it is generally recognised that recourse to proxy variables is possible if information on a characteristic is not accessible or if accessing it would be unduly costly or time-consuming. Recourse to proxy variables must, however, be well justified, the proxy variable must be highly correlated with the missing consumer attribute and the limits on the information value of the overall model that result from the use of proxy variables must be made transparent. On the use of proxy variables in the context of credit scoring, see Berg, Burg, Gombović and Puri, 2018).

³⁶ Hamburg Local Court judgment of 16 March 2017 – case reference 233 OWi 12/17. See the Heise Online report at <https://www.heise.de/newsticker/meldung/Datenschutzverstoss-15-000-Euro-Bussgeld-wegen-Geoscoring-3664654.html>, accessed on 15 May 2018

³⁷ <https://www.sparkassen-direkt.de/telematik/faq/>; accessed on 24 May 2018

If, on the other hand, proxy variables are used to predict individual consumer behaviour – with considerable economic implications for consumers in some cases – the need to justify the use of such variables for scoring purposes is far greater. Given the individuality of the scored person, the risk of the complete misjudgement known as the ecological fallacy (Kamp and Weichert, 2005) is greatest in these circumstances. This fallacy entails wrongly deducing individual data from aggregated ('ecological') data. From information about living conditions in the neighbourhood of a scored person, conclusions are drawn about that person's financial situation in general and about his likelihood of defaulting on a loan in particular (Kamp and Weichert, 2005). From a correlation that appears to exist within the population as a whole, a causal relationship is inferred for an individual.

This is acceptable and expedient from the point of view of a business that seeks to avoid payment defaults and is not perturbed by the loss of turnover but not in the eyes of an individual who is wrongly assessed. In the case of geo-scoring, the ecological inference fallacy makes consumers jointly responsible for their neighbours' misconduct and curtails their sovereignty. It follows that a person cannot improve his score by gradually altering his own behaviour but can only do so by taking an invasive measure such as moving to a 'better' neighbourhood, that is to say an area where average solvency ratings are higher.

3. Weighting of input variables

The score that is assigned to a consumer depends essentially on the nature of the individual data and their relative weighting. In an extreme case, this may mean that consumers who tick most of the boxes for which points are awarded nevertheless receive a lower overall score because disproportionately more weight is attached to other data items. Three different weighting systems can be distinguished:

Weighting on the basis of regression parameters: In an algorithmic decision-making process, the weighting of a consumer attribute is determined by the influence of that predictor variable on the target variable; the weighting is often defined in the form of parameters in regression equations but can also be defined in other ways (see the background note above as well as Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). However, since providers of credit scores regard the weighting of consumer attributes as part of their trade secret, a view that the Federal Court of Justice has endorsed, there are scarcely any research findings on how the weighting is determined and whether it is objective. There are no specific statutory provisions governing the weighting of predictor variables.

Heuristic methods used in corporate practice: In the realm of telematics-based motor insurance – for example in the S-Drive tariff offered by the insurer Sparkassen DirektVersicherung – a rule of thumb (heuristic approach) is used whereby the consumer is scored on the basis of the attributes driving style (acceleration and braking), speeding, night driving and urban driving, which are weighted very neatly at 40%, 30%, 20% and 10% respectively.³⁸ It is impossible to say whether and to what extent these weightings reflect the relative influence of each of the listed consumer attributes on the target variable.

38 Abgerufen am 23. Mai 2018 von URL <https://www.sparkassen-direkt.de/telematik/faq/>.

Model-independent weighting: A familiar feature of bonus programmes is that the number of available bonus points is not necessarily determined by the actual beneficial effect of a health-promoting activity (consumer attribute) on a consumer's health (target variable), i. e. is not necessarily model-dependent, but may relate to the time and/or expense that the consumer devotes to the activity (model-independent weighting). Since there are scarcely any research findings on model-independent weighting, in-depth discussion of this issue is required, which we shall undertake in Part C. The link, i. e. the statistical correlation, between a predictor variable and the target variable is therefore largely severed in the case of these weighting factors.

Adverse effects on consumers may result from the weighting of attributes in the following cases: (1) if the applied weighting factors are not clear to the competent supervisory authority, or possibly even to consumers, and the impact of consumers' own behaviour on their score is unforeseeable; (2) if weighting factors vary considerably between scorers, in other words if a consumer attribute to which one score provider attaches a great deal of weight is irrelevant to another provider, a situation that is most likely to occur if weighting factors are not based on objectifiable criteria such as the value of the regression parameters; (3) if weighting factors change over the course of time.

VI. Competing fairness criteria

Another scoring problem arises in connection with the specific composition of data sets that are used to determine a score. In the realm of machine learning, this kind of data set that is used for statistical analyses is known as a training data set (details on the principles of machine learning are presented in chapters IV.1 and IV.2 and in *Gesellschaft für Informatik*, 2018, section 4.1). The important thing in this context is that specific elements of the baseline data impact directly on the structure and predictive power of the model. For example, in a data set relating to creditworthiness, three quarters of all persons deemed creditworthy may be female and only one quarter male. In reality – in this case after adjustment for ‘other’ genders – the balance between the sexes is about 50/50, and so we have here an imbalance in the baseline data from which any statistical software (‘learning algorithm’) is very likely to infer that attributes other than gender play only a minor part and that the probability of creditworthiness depends primarily on the sex of the applicant (*Gesellschaft für Informatik*, 2018). This phenomenon is sometimes referred to in literature as bias amplification (Zhao, Wang, Yatskar, Ordonez and Chang, 2017; cf. *Gesellschaft für Informatik*, 2018).

Fairness in the calculation of scores is a fundamental problem that arises irrespective of the method used in practice – even rules of thumb and other heuristic methods can be unfair. The literature on fair machine learning sheds the best light on these fundamental problems, discussing quantitative methods designed to guarantee the most comprehensive possible equal treatment of individual groups and persons (Dwork, Hardt, Pitassi, Reingold and Zemel, 2012; *Gesellschaft für Informatik*, 2018; Kleinberg, Mullainathan and Raghavan, 2016). The key conceptual terms in this discussion are overall accuracy equality, statistical parity, conditional procedure accuracy equality, conditional use accuracy equality and treatment equality.

Berk *et al.* (2017) summarise the combination of all five aspects of algorithmic fairness in the concept of ‘total fairness’ (cf. *Gesellschaft für Informatik*, 2018). It is most likely impossible, however, to achieve total fairness by adjusting algorithms, because the various fairness criteria are in competition with each other in the sense that all of the fairness criteria could never be fulfilled simultaneously. This conclusion is reached by Chouldechova (2017) and Kleinberg *et al.* (2016), who analyse three measures of fairness and show that no method currently in existence meets all three quantitative fairness criteria simultaneously. It is highly probable, because of the prevalence of different risks among different groups, that there will never be a method that can achieve all fairness criteria at the same time.

Competing measures of fairness: a numerical example³⁹

Let us assume that there is a wide variation in the actual risk of payment default between two groups, say low earners and high earners, whose respective default probability rates are 5% and 0.5%. Assuming that a score has the same predictive power of 90% accuracy for both groups, the number of right and wrong predictions per group will be as follows if each group comprises 10,000 persons:

In the high-risk group there will be 500 payment defaults (5% of 10,000), of which 450 will have been predicted by the score and 50 will have been missed. Of the 9,500 cases in which there is no default (95% of 10,000), 8,550 will have been correctly predicted thanks to the predictive power of 90%, but 950 will have been wrongly marked as likely defaulters. The percentage of correctly predicted payment defaults for the high-risk group will therefore be $450 \div (450 + 950) = 32\%$.

For the low-risk group there will be 50 payment defaults (0.5% of 10,000), of which 45 will have been predicted by the score and 5 will have been missed. Of the 9,950 cases in which there is no default (99.5% of 10,000), 8,995 will have been correctly predicted thanks to the predictive power of 90%, but 955 will have been wrongly marked as likely defaulters. The percentage of correctly predicted payment defaults for the low-risk group will therefore be $45 \div (45 + 955) = 4\%$.

The difference in the percentage of correct predictions stems from the fact that, where the risk is minimal, very many non-risky cases are wrongly classified (false positives). The percentage of correct default predictions would only be roughly equal for both groups if the algorithm for the low-risk group were ten times more accurate than that for the high-risk group, in other words if the predictive power of the algorithms were about 99% and 90% respectively. This is a very unlikely scenario and does not generally occur.

The consequence for our example is that, if both groups are treated equally in terms of accuracy, i.e. specificity (conditional procedure accuracy equality), the low-risk group runs a far higher risk of false positive assessment (conditional use accuracy equality).

The achievement of quantitative fairness also has a side-effect: if all consumer attributes which have a statistically significant effect on the target variable are selected – which modern machine-learning processes do automatically – these may include discriminatory and hence legally protected grounds or closely associated attributes. If discriminatory grounds are eliminated from the statistical model on that account, the statistical model as a whole will become more imprecise. The more attributes that are removed because of their association with membership of a particular group, the more precision will be lost, and the quality of the statistical model will suffer accordingly (*Gesellschaft für Informatik*, 2018). In short, this creates an irresolvable conflict of aims between avoidance of recourse to protected grounds, even though they may be significantly influential, and the quality of the score. If statistically significant variables are not used, more people will be inaccurately scored. It follows in turn that we are confronted with conflicting fairness criteria which, in general, cannot be simultaneously achieved. An optimum solution must be sought on the basis of fairness priorities.

Which measures of fairness are to be prioritised and which are to be subordinated cannot be decided by mathematical formulae and machine learning. What is needed is social accord on the legitimate purposes and uses of attributes.

³⁹ Cf. also the definitions and specimen calculations in *Gesellschaft für Informatik*, 2018, sections 4.3.1 and 4.3.2.

VII. Consumers and society: expectations, knowledge, competence and implications

This chapter provides an overview of the state of research into consumer expectations and acceptance of scoring and into consumers' scoring-related knowledge and digital literacy in Germany. It should be said from the outset that scarcely any independent academic studies in Germany have examined consumers' knowledge and digital literacy regarding established scoring systems, such as credit scoring, and potentially novel systems (in areas such as healthcare or in the calculation of composite scores from various areas) as well as the associated implications (for exceptions, see Fischer and Petersen, 2018, although that work relates to algorithmic decisions in general, Müller-Peters and Wagner, 2017, and PricewaterhouseCoopers, 2018). For this reason, the SVRV commissioned a representative survey (see Part D), one of the aims of which was to form a picture of knowledge and acceptance of scoring in various areas of life among the resident population of Germany.

1. Consumers' expectations and acceptance of scoring

If consumer policy relating to scoring is to be shaped in such a way as to focus on the justified and socially legitimate expectations of consumers, the first task will be to identify these expectations. Credit scoring has a long tradition, but scoring practices in other areas, such as telematics-based motor insurance and health insurance, are relatively new phenomena which are only gradually coming to play a part in various aspects of consumers' lives; accordingly, independent and informative academic studies that shed light on consumers' attitudes and expectations relating to scoring are still a rarity in Germany. It therefore seems advisable to establish empirically in which areas, to what extent and in what form the scoring of consumers is regarded in Germany as legitimate and in what cases it is held to be unwarranted.

For example, which attributes do consumers regard as legitimate and justified predictor variables for the assessment of their creditworthiness or of their motor or health insurance premiums and which do they not? An-

other question that arises is whether and to what extent consumers tend to approve or disapprove the linking of their scores and predictor attributes from various areas of activity. Another issue entirely is whether society as a whole shares the consumers' appraisal of what is warranted and legitimate. In a democratic society, finding this out is ultimately incumbent on the parliamentary legislature, which must translate these moral and ethical perceptions into statutory rules or else decide to refrain from regulatory intervention.

With regard to traditional credit scoring, acceptance of the communicated scores clearly seems to be relatively low. In a representative study, for example, more than half to three quarters of the respondents considered their score to be unfair, although the level of acceptance of scores depends on the company from which consumers obtained their personal credit records (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). In general terms, acceptance is greater where scores are higher, i.e. more favourable, but not even high scores are necessarily perceived as fair. We can only speculate on the reasons for this situation. In the same survey, for instance, almost half of the respondents stated that they found the explanations given by credit reference agencies to be inadequate and often incomprehensible. More than 80% of the respondents, moreover, wished for more transparency and information from credit reference agencies and supported an information access obligation for those agencies (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014).

Social scoring, a novel method used to determine creditworthiness in which data are obtained from social networks, was assessed by more than half of the respondents (56% of a sample numbering 1,023) as risky in a representative study dated 2018 (PricewaterhouseCoopers, 2018). The majority of respondents (71%) stated that they saw the danger of flawed conclusions being drawn in credit reports as a result of the use of data from social networks. More than half of the respondents in the 18–25 age group, however, said that they would favour social scoring if certain transparency criteria were met, such as disclosure of the calculated score and information on the data used for scoring purposes (PricewaterhouseCoopers, 2018).

Germany's National Academy of Science and Engineering (acatech) had a representative survey conducted at the end of 2017 (sample size 2,002) on the subject of technology with the main focus on digitisation (acatech and Körber-Stiftung, 2018). In general, digital technology was viewed rather sceptically by the majority of respondents. In the case of autonomous driving, it emerges that the main concern expressed by the majority of respondents relates to data security. In addition, the collection of personal data by the vehicle received a disapproval rating from almost two thirds of the respondents. Since telematics-based insurance tariffs involve the recording of comparable data, their acceptance is also in doubt.

A widespread uneasiness regarding the disclosure of personal data seems to indicate an underlying scepticism about scoring-based business models founded on big-data analyses. For example, two thirds of the respondents in Germany fear that companies are collecting excessive personal data through the Internet. The greatest concerns, each echoed by almost 80% of the respondents, relate to the buying and selling of personal information, to a general lack of protection of personal data and to the danger that personal information could come under surveillance (Centre for International Governance Innovation, 2017). It is not easily fathomable, however, why so many consumers nevertheless consent quite readily to the storage and processing of their data. The offices of Data Protection Commissioners are not by any means inundated with thousands of complaints from consumers putting up resistance against being half-coerced into consent. There is a considerable divergence between political ideals and practical action.

Consumer acceptance and expectations of telematics-based motor insurance and lifestyle-based health insurance, including potential future developments in these fields, was the subject of a survey conducted on behalf of Cologne University of Applied Sciences (TH Köln).⁴⁰ Almost half of the respondents (46% of a sample of 834) could imagine having data on their driving behaviour recorded and passed on to their motor insur-

er as a means of having their premiums considerably reduced. In addition, basing premiums on modifiable attributes such as careful driving tends to be accepted by the majority of respondents, whereas factors that drivers cannot influence through their driving behaviour, such as whether their driving is done during the day or at night, tend to be rejected as pricing variables by the majority. In the case of health insurance, between half and two thirds of respondents consider it fair that modifiable behavioural criteria such as whether a person attends screening examinations or smokes or drinks to excess should be considered when premiums are set. The vast majority, however, believe that attributes which cannot be changed, such as a family history of particular medical conditions, should not be factored into the calculation of insurance premiums. In principle, more than a third of respondents would sign up for a lifestyle-based health-insurance tariff if it would save them money (Müller-Peters and Wagner, 2017).

Consumers recognisably tend to be prepared in principle to make personal behavioural data available to insurers, particularly if it can obtain them a price cut (Müller-Peters and Wagner, 2017). Which attributes consumers regard as acceptable and unacceptable for consideration in the calculation of insurance premiums appears to depend on the type of insurance (motor or health insurance) and the modifiability of the attributes (e.g. family medical history versus alcohol consumption).

It has so far remained a moot point, however, whether individual factors such as being personally affected (e.g. one's state of health), socio-economic status, demographic variables and specific attitudes to things like data privacy and technology as well as one's locus of control alter consumers' attitude to and acceptance of scoring. If a high-definition image of consumer expectations and acceptance of scoring is to be obtained so that tailor-made measures of consumer policy can be adopted, it is important that specific consumer categories be identified.

⁴⁰ The survey was conducted by an institute closely associated with the insurance industry; the findings, however, are essentially comparable with our own survey (see Part D).

Insurers who make initial forays into scoring often advertise only beneficial implications, that is to say a bonus system whereby policyholders, by behaving in particular ways, can collect points with the prospect that a certain number of points will qualify them for material rewards or a reduction in their insurance premiums.

The opposite scenario, that particular behaviour is liable to have adverse consequences, in other words a system of penalty points whereby a policyholder's conduct may, for example, increase his or her insurance premiums, has not yet been incorporated into the building blocks of a telematics-based insurance tariff. Frequency of accidents has never yet been one of the variables that are used in calculating a person's score. The bonus programmes offered by the statutory health insurance scheme also entail only bonuses, and non-participation in measures does not result in penalties such as dearer insurance premiums, and this indeed is in line with the relevant legal provision, section 65a(3) of Book V of the German Social Code. It therefore seems logical that consumers' attitudes to and acceptance of behavioural premiums will also vary in accordance with the prospective consequences. If a telematics-based motor insurance tariff offers only benefits, such as lower insurance premiums for adherence to statutory speed limits, its acceptance level will presumably differ from that of a tariff which also involves financial penalties, i. e. as well as not awarding bonus points to drivers who exceed the statutory speed limit, the insurer also penalises them by charging them more. For this reason, another objective of the representative public survey commissioned by the SVRV was to establish the extent to which the acceptance or rejection of a behavioural pricing system for motor and health insurance premiums would be affected if the system involved both bonus and penalty points (see Part D below).

2. Knowledge and competence

2.1. Consumers and algorithms: knowledge and attitudes

According to a recent survey on a representative sample of 1,221 persons by the Bertelsmann Foundation (Fischer and Petersen, 2018), a lack of knowledge about algorithms and ambivalence or reservations about their use are prevalent among most of the resident population of Germany. Since scoring models are based on algorithms as a rule, excerpts from the findings of that survey are presented in the following paragraphs.

Although three out of four respondents in the study say that they have heard the term 'algorithm', almost half of the sample cannot describe spontaneously what it means. Of those who have heard at least once of algorithms, however, more than half know nothing of how algorithms basically work. Only one tenth of these respondents claim to know how algorithms, as they understand the term, actually function. Whether the respondents associate anything with the term 'algorithm' and know how algorithms work varies widely with age, education level and gender: respondents with an *Abitur*, the German university entrance qualification, are far more frequently able to express at least a vague understanding of algorithms than those with lower school qualifications, male respondents more frequently than female respondents and persons under the age of 45 more frequently than the over-60s. The extent to which respondents are aware of the use of algorithms also varies from one area of activity to another: whereas more than half of the respondents were aware – or claimed to be aware – that algorithms were used in individualised advertising on the Internet, and half, or rather just less than half, were aware that algorithms are used in facial recognition in the context of video surveillance and in the assessment of creditworthiness, only about a third are aware that algorithms are used in some regions to analyse staffing requirements and for police operations, in which they identify areas where the risk of burglaries is particularly high (predictive policing). And slightly fewer than one fifth of respondents were aware that algorithms can also be used by the judicial authorities to assess the probability of reoffending.

In the view of more than one third of the respondents in the Bertelsmann study, the risks inherent in algorithm-based decisions outweigh the opportunities they offer, whereas fewer than a fifth see them primarily as an opportunity (Fischer and Petersen, 2018). Almost half of the respondents are undecided as to whether risks or opportunities are preponderant. This suggests that a large percentage of the resident population of Germany has not yet formed a clear opinion on this matter. This should not come as a surprise, since there is no clear evidence yet of the actual risk-benefit ratio for many algorithm-based decisions, and even among experts the question remains a source of controversy.

A firm opinion among respondents is recognisable when it comes to the question whether decisions should, as a matter of principle, be made by algorithms or by humans. A very large majority of respondents (79% of the sample of 1,221) feel uncomfortable with algorithmic decisions and prefer human decisions. Broken down into areas of activity, a more differentiated picture emerges: in spite of an overwhelming general rejection of exclusively algorithm-based decisions, a majority would consent to the decision on efficient use and administration of storage spaces being left to algorithms. Almost half, moreover, would be in favour of exclusively algorithmic decisions on individualised online advertising and spellchecking in the field of word processing. Most of the respondents, on the other hand, believe that the assessment of creditworthiness, medical diagnoses and identification of the probability of re-offending should be undertaken exclusively by humans or at most by humans taking decisions with the aid of algorithms. The final decision, they believe, should be taken by a person. In short, particularly in sensitive areas such as creditworthiness, criminal justice and health, the majority oppose the use, or at least the exclusive use, of algorithm-based decisions.

In line with a predominantly unfavourable attitude to algorithms, almost two thirds of respondents across the whole education and age spectrum support tighter control of algorithms. Measures designed to control the use of algorithms such as compulsory indication of algorithmic decisions, disclosure of algorithms to independent experts and the introduction of an ethics commission meet with the approval of the overwhelming majority of respondents (Fischer and Petersen, 2018).

In general terms, the findings of the representative survey described above seem to indicate that there are currently wide gaps in Germany in people's knowledge of what an algorithm is and how it works. The majority of the population have scarcely looked into the subject of algorithms, which ties in with the fact that only a minority have a definite opinion on algorithms in general. At the same time, in many areas of activity decisions assisted by or exclusively based on algorithms meet with a great deal of scepticism and rejection. Accordingly, if algorithms in their various fields of application are to be better understood and their pros and cons more objectively assessed, it seems logical to pursue the aims of reducing the knowledge deficit and developing digital literacy among the resident population of Germany.

In addition, a balanced social debate should be initiated on the demonstrable implications of algorithms with a view to addressing fears, rejection and challenges. Equally, however, there is a need for education about the empirically substantiated potential and opportunities offered by new technologies and algorithms and hence by scoring too.

2.2 Consumers' knowledge of scoring-related matters

Knowledge relating to scoring tends to be minimal among consumers. Many US citizens, for example, are not familiar with their personal credit rating. One tenth of respondents in a survey described in Levinger, Benton and Meier (2011) substantially overestimated their own score and hence their creditworthiness. Moreover, many consumers in the United States find it difficult to reconstruct the details of their credit record. Younger consumers in the 18–34 age brackets know particularly little about the details and implications of their credit scores (Consumer Federation of America and Vantage Score Solutions, 2016). This is particularly relevant, because a field experiment in the United States demonstrated that individuals' knowledge of their credit score was linked with fewer payment delays and, in turn, with an increase in their credit score over the course of a year (Homonoff, O'Brien and Sussman, 2017).

Not only in Germany do consumers often have no knowledge of their information rights regarding their credit rating. This is not too surprising in view of the fact that consumers' overall awareness of their rights is low. For example, more than half of the respondents in a representative survey dating from 2014 did not know that they were entitled to obtain, free of charge, a copy of their own personal credit record (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014). What is more, consumers are often unaware that the information about them which is stored by credit reference agencies may be incorrect or outdated. This happens, for instance, when a bank omits to notify a credit reference agency that a loan has been fully redeemed. This omission may adversely affect the borrower's credit score in a specific case as well as whether and on what terms the borrower can obtain any subsequent loan. Many consumers are also unaware that they are entitled to have their score corrected or where and how they can exercise that right. Most consumers, moreover, do not know exactly which attributes credit reference agencies use to calculate their credit score (see Part D below).

2.3 Competence in the context of scoring

Digital literacy is a major prerequisite for responsible and informed decisions. This applies not only in the digital world in general but also with regard to scoring in particular. Although the digital world influences the everyday lives of consumers, digital literacy certainly cannot be taken for granted but must be supported and fostered by policymakers. This point has already been highlighted by Advisory Council for Consumer Affairs in its report on *digital sovereignty* (SVRV, 2017a). According to that report, digital literacy, along with technology and consumer-friendly regulation, is a key component of the 'digital sovereignty triangle'. Appropriate education, application of knowledge, ability to self-monitor and willingness to engage in lifelong learning are essential keys to the attainment of digital literacy (SVRV, 2017a).

Competent and critical assimilation of scoring and its implications, however, also depends on consumers acquiring digital risk intelligence, which is a specific form of risk intelligence. This means obtaining the knowledge, know-how and skills that are the key to critical understanding and reflective handling of insecurity problems in the digital world. As is suggested by the representative surveys cited above, these assets are, in many respects, non-existent. Informed decisions, in which consumers are supposed to be able to weigh up potential pros and cons, are therefore thrown into question. Some consumers are liberal with their personal data if that earns them advantages like discounts or bonus points (Geslevich-Packin and Lev-Aretz, 2016) and are presumably not very keenly aware of possible adverse consequences. It seems that only a relatively small number of consumers are fully conscious of this balancing exercise; many consumers, moreover, appear to be unaware that companies use sensitive information – which is often willingly disclosed to them by consumers – for their own purposes and possibly against the consumers' interests (for a summary treatment, see Geslevich-Packin & Lev-Aretz, 2016).

The competence of a consumer in the specific context of scoring rests on the following foundations: scoring-related knowledge, for example knowledge of the variables that are used to calculate the score in question, knowledge of the utilisation and purpose of scoring (when, where and why it is used) and knowledge of the consequences arising from scoring but also practical knowledge of where and how to obtain information about one's own score. In practice, a competent participant in a scoring system would be capable of critically examining and assessing its quality, of at least questioning its discrimination potential and of finding and using alternatives if any were available.

Unless consumers have sufficient knowledge and the competence to understand and question scoring-based decision-making processes, their self-determination and self-monitoring will be limited. It seems essential that consumers should be able to understand and gauge for themselves the potential benefits and damage deriving from their own actions, such as the disclosure of their personal data.

Another relevant and important aspect of our approach to scoring is data literacy, in other words the capacity to handle data methodically, to use them in a given context and to question them. This includes the ability to collect, administer, analyse, interpret and apply data (Ridsdale *et al.*, 2015).

Last but not least, in the absence of the conditions described above, a social debate on scoring would scarcely be possible. For this reason, part of the present report is devoted to the pre-existing stock of scoring-related knowledge and competence among the population.

3. Social implications

The impact of consumer scoring is not felt by individuals alone, for example through behavioural insurance premiums. On the contrary, such scoring can impact on the whole of society at many levels. Moreover, it raises ethical and moral questions in the field of tension between individualisation and solidarity. Lastly, it also provides food for debates on fundamental issues, such as the extent to which basic values of a democratic society like personal freedom, autonomy and solidarity are challenged by scoring. As we have outlined in chapter B.II above, there is also a need to examine whether scoring contributes to social inequality or helps to reduce it and whether it systematically disadvantages or unduly favours particular groups of consumers. These issues cannot be addressed in detail in this report but can only be outlined. At the present time, many of the described implications of the growth of scoring for German society are still speculative.

3.1 Social functions

In principle, certain forms of scoring promote the functioning of the economy and therefore serve the legitimate interests of society. Traditional credit scoring protects businesses from payment defaults, which makes it an essential basis for lending to consumers. Since defaulters tend to be filtered out, reliable consumers receive more favourable terms than would be possible in the absence of scoring. Scoring-based decisions may also help to ensure that consumers are subject to fewer arbitrary decisions based on the conscious or unconscious prejudices or misconceptions of exclusively human decision-makers. For consumers and businesses, moreover, it is easier to enter into long-term contractual arrangements, such as mobile-phone contracts, if there is recourse to an essentially reliable system of creditworthiness and solvency checking (Schröder and Taeger, 2014).

Behavioural tariffs, on the other hand, may be assessed in fundamentally different ways when viewed from a normative and ethical perspective – one example being the difference between third-party motor insurance and health insurance (see section E.II.2 below). Both private health insurance and third-party motor insurance are based on a system of premium assessment. In each case the insurance

premium is individually tailored to the risk profile (see section E.II.2). In the case of private health insurance, a person's medical history and age are relevant factors, while in motor insurance no-claims periods and place of residence play a part. Every policyholder, then, is charged on the basis of his risk profile. Some motor insurers have already introduced telematics-based tariffs in which premiums are based on attributes that include aspects of driving behaviour, such as recorded driving speeds. According to insurers, scores based on driving behaviour are associated with a lower risk of accidents. Policyholders with a better score pay less, the premium being adapted not only on the basis of driving behaviour itself but also on the basis of situational factors, such as the predominance of rural or urban driving. In this context, telematics-based scoring can have a socially beneficial impact by reducing the cost of medical treatment arising from accidents. Should telematics-based tariffs actually reduce the risk of accidents, then, their benefits would not only accrue to individuals but to society in general. Such benefits, however, have yet to be demonstrated by relevant academic studies.

3.2 Implications for solidarity

Scoring and the associated consequences may be perceived as fair towards the individual consumer, whereas scoring may prove to be unfair to society as a whole and possibly contrary to the principle of solidarity, and vice versa. For example, is it not fair to an individual policyholder if the amount of his insurance premiums is determined by his voluntary behaviour, in other words if scores depend on how health-consciously people live their lives or how carefully they drive? Is it not anti-social to speed and so potentially endanger other road users or, in the event of an accident, burden the health system with the cost of the associated medical treatment? As far as the individual is concerned, people might also wonder, for example, why they should contribute to cover the risk of illness for a neighbour who is a smoker when they themselves have a health-conscious lifestyle. Conversely, there is sound logic behind the argument that it is a social value to show solidarity with our neighbours' freedom to choose their own lifestyle, even if it means sharing the healthcare costs that arise from smoking. There again, reasons of privacy preference may also warrant an active decision to opt out of being scored.

What is regarded, in the context of scoring, as cooperative and fair or ethically and morally imperative must be considered and carefully weighed up in a social debate. The protection of individuals and/or specific groups reflects the general interest of a society based on solidarity. It must therefore involve harmonising the interests of individuals or particular groups with the legitimate interests of society. This also means that particularly vulnerable persons or groups must not be systematically disadvantaged as a result of scoring, for example through an above-average incidence of false positive scores. The social debate would have to determine which values and standards are shared and how the interests of individuals or especially vulnerable groups can be reconciled with the legitimate interests of society as a whole.

The question whether scoring might undermine solidarity within society in general and the system of statutory health insurance (see section E.II.3 below) in particular is already the subject of lively debate (see, for example, Deutscher Ethikrat (German Ethics Council), 2017). From a social perspective, solidarity may be understood as a pre-eminent social asset of a democratic society, an asset based on shared values and standards. It is a consensus view, for instance, that specific groups such as vulnerable consumers, e.g. the socially underprivileged and people with disabilities or infirmities, who are in particular need of help must be protected, integrated into society and given social and, where appropriate, financial assistance (Micklitz, Oehler, Piorkowsky, Reisch and Strünck, 2010). In the realm of statutory health insurance, the solidarity principle means that insurance contributions are not based on a person's health status or lifestyle but on each insured person's individual income. Those in good health exercise solidarity by picking up the bill for insured persons in a poorer state of health, even though their poorer health may be due to voluntary decisions, such as the adoption of a particular lifestyle.

All insured persons are entitled to the same benefits, regardless of the size of the contributions they pay. A core principle of the risk-sharing community on which the statutory health insurance system is founded is solidarity between the healthy and the sick. The basis of this solidarity is a sufficiently large number of healthy contributors compared with relatively few contributors with illnesses and therefore greater demands on health services.

Since health care is a particularly sensitive area and merits special protection, it follows that the statutory health insurance system is highly regulated. Accordingly, lifestyle-based insurance tariffs have not yet been introduced into the statutory system in Germany. Lifestyle-based tariffs would, for example, systematically disadvantage people who are unable to adopt a particular mode of behaviour or who are not prepared to communicate physical data, some of which may be sensitive, to their insurer. If, for example, walking a certain number of paces daily were a factor in the pricing of insurance premiums, people with mobility problems, such as someone with a broken leg, would be at a considerable disadvantage, because their score would be lower, and in some circumstances they might have to pay more. In general terms, then, there is a danger that lifestyle-based tariffs might undermine the principle of risk-sharing solidarity. In the realm of statutory health insurance, the bonus programmes that already exist, within which certain people may be disadvantaged, are proving problematic in this respect. The difficulties are compounded by the practice of crediting measures for which the risk-benefit ratio for individuals is so disputed, as in the case of the prostate-specific antigen (PSA) test for men aged 50 and over (see Ilic *et al.*, 2018), that their coverage by health policies has not been authorised.

3.3 Implications for the social structure

The potential consequences of scoring may have social implications. An increase in self-monitoring, for example with regard to sleep patterns, exercise and diet, by means of smart devices, and in mutual assessment, for instance through social networks, may generate increasing pressure on people to do more and more to better themselves in accordance with the ideals of a performance-driven society. This could lead to a split into optimised and suboptimised individuals (Selke, 2014). It is thus conceivable, although it still seems a long way off, that social status will one day be measured not only by people's formal education, occupation and income but also by the number of recorded paces they walk each day, their general health score or their driving score. Even today, in some areas of people's lives their status and value are already being described, created and fixed by scores – in short “Numbers make people”, as Steffen Mau (2017) phrased it succinctly in his book *The Metric Society – On the Quantification of the Social*. For example,

the Klout website, which stopped providing its services at the end of May 2018, scored people on a scale from 1 to 100 points on the basis of data from social networks, the purpose of the Klout score being to reflect each individual's online social influence.

It is already common practice to categorise people into specific target groups on the basis of scoring so that consumers can then be addressed and treated in different ways. Such classification is not merely descriptive and its aim is not simply to define differences. In fact, it can also imply a social selection between ‘valuable’ and ‘less valuable’ consumers (Selke, 2014). There is consequently a danger that consumers from the lower categories may find it difficult in some cases to obtain, for example, a mobile-phone contract, motor insurance or health insurance on anything approaching reasonable terms. This raises another question, namely how easy or difficult would it be to move up from a lower to a higher category (Mau, 2017)?

Through the increasing collection and correlation of data from various areas of activity, e.g. leisure pursuits, consumption patterns, payment history, health status, exercise habits, membership of social networks, area of residence, working life, occupation and family status, for the purpose of consumer categorisation (Saetnan, Schneider and Green, 2018), it certainly appears conceivable that scoring could form the basis for the emergence and establishment of a digital class society. In this scenario, individuals would be locked in permanent competition for a good score. The potential this offers for an increasing individualisation of society, accompanied by diminishing solidarity, could lead to a weakening of social cohesion. If, as appears to be on the horizon in China, even ‘friends’ on social networks are chosen on the basis of their scores (Kostka, 2018) because having friends with low scores could impact on a person's own score, new forms of social exclusion may conceivably emerge. In this context, the social pressure to be scored and to adopt ‘score-compliant’ behaviour must not be underestimated.

3.4 Implikation bezüglich der Wahlfreiheit

As regards telematics-based tariffs, it could be an issue in some cases, once these tariffs become sufficiently widespread, that more and more of the population could be induced or feel themselves compelled to choose such policies. It is conceivable, for instance, that non-participation and non-disclosure of personal data and other information might be interpreted as potentially indicative of a higher risk and could lead to stigmatisation and monetary penalisation. Since people with the prerequisites for a good score will tend to be more inclined to disclose data and information about themselves for scoring purposes, it is conceivable that, in some circumstances, individuals will also feel compelled to disclose the data and information that are required for participation in a telematics-based insurance scheme in order to avoid any disadvantages that might result from non-participation. Peppet (2011) refers to this effect as the unravelling of privacy.

Back in the 1920s and 1930s, the Hawthorne effect was described; in simplified terms, it is the phenomenon whereby people alter or adapt their behaviour when they know that they are being observed (French, 1953). This effect could also be a factor in the context of scoring. Society will therefore face the challenge of preserving the pre-eminence of democratic values such as personal freedom, including freedom of choice, as well as privacy, autonomy and solidarity. Preservation of these values also implies, for example, choosing a supposedly unwise course of action, not adapting one's behaviour to the wishes of society and actively opting out of being scored. The foregoing social implications of scoring in Germany must still be regarded as speculative at the present time. It is, however, highly probable that scoring will impact on society, though in what form and to what extent remain to be seen. That is precisely why it is important to create awareness of the possible social consequences of scoring. In this context there is a particular need to discuss which ethical and moral values are indispensable for a free democratic society, which of the implications of scoring should be regarded as socially unacceptable and which are to be deemed socially acceptable.

VIII. The danger of a super score

Processes that use big data or machine-learning algorithms and undertake consumer scoring on the basis of huge volumes of data and complex algorithmic calculations have not yet been used in Germany in the sectors under examination in this report (see Part C below). Outside Germany, however, a trend towards such business models is observable in various areas, as this chapter will show. This development should also be seen in connection with the scope for purchasing data from specialised data traders. Large volumes of data generated by data trading could potentially make it possible to repersonalise anonymised or pseudonymised data. The correlation of personal data from various areas of people's lives and from various sources also means that the creation of super scores, similar to the Social Credit Score in China, is a danger which cannot be ruled out.

By 2020, the Chinese Government is planning to assess the conduct of all citizens and businesses by means of a system of social credit scoring. 'Trustworthy' behaviour as defined by the Communist Party is to be rewarded and 'untrustworthy' behaviour punished. This can have implications for people's training and career as well as for every aspect of their everyday lives.⁴¹ One of the official aims of the system is to increase trust between market participants. Initial effects on China's foreign partners and customers were also observable (Hoffmann, 2018). There are currently pilot projects in 40 Chinese cities. Some commercial providers also offer social credit systems on a voluntary basis. The processes are based on algorithmic big-data analyses. A study on public perception of such systems has shown that some 80% of China's Internet users rate the governmental and commercial social credit systems in their countries favourably (Kostka, 2018).

The question arises whether similar scoring systems are conceivable in Germany too (on this point see also Al-Ani, 2008). Although it is not to be expected that governments in the Western world will initiate such developments, companies with this kind of business model could pursue the aim of aggregating data from various

areas of people's lives and using them to calculate super scores with the aid of algorithms.

In the following paragraphs we shall describe some examples of big-data-based scoring models outside Germany as well as addressing the issue of data trading. In so doing, we intend to highlight current opportunities and trends in the realm of scoring that could eventually become relevant in Germany too. If they do, it may also highlight the great importance of the relatively high level of regulation in Germany.

The scenario of a super score – which, in certain circumstances, could even be the product of consumers' freely given consent to make their data available for transfer to third parties – is explained below.

1. Scoring models abroad

Credit scoring

More and more businesses in the financial sector are developing scoring processes for the purpose of predicting people's creditworthiness by means of big data (Jentzsch, 2016). Use is also made of unconventional data, that is to say data without any direct link to financial creditworthiness, such as activities on social networks or online browsing and purchase histories. Target groups are mostly the so called 'underbanked', in other words people who, for reasons such as the lack of a conventional financial history or the absence of credit standing with traditional credit institutions, would not be able to obtain a loan. This approach is chiefly designed to open up markets in developing or newly industrialised countries.

⁴¹ It is known, for example, that people's score may influence whether they are allowed to buy flight or train tickets or send their children to a private school. Those with high scores, moreover, are eligible for preferential treatment in hospitals and are exempted from the payment of car-club deposits (Kostka, 2018). In 2014, Sesame Credit Management, a subsidiary of Alibaba, announced a cooperative venture with the Luxembourg Consulate-General in Shanghai, whereby Chinese nationals with a high Sesame score can submit simplified visa applications through an online gateway (Alibaba Group, 2014).

The Kreditech company⁴² was founded as a start-up business in 2012 and has its headquarters in Hamburg (Gründerszene.de, 2018). Through subsidiaries, Kreditech currently processes loan applications in Russia, Poland, Spain, the Czech Republic and Mexico.⁴³ Unlike traditional credit reference agencies, Kreditech does not look primarily for past defaults in order to establish a person's creditworthiness but evidently uses proxy variables and assesses big data, according to its own publicity, by means of artificial intelligence and algorithms, which are constantly fed with additional data and hence somewhat misleadingly referred to in publicity material as 'self-learning' (Seibel, 2015).

According to the company, its software identifies customers online within a few seconds and performs a fully automated calculation of their credit score. If an application is approved, the money is transferred to the customer's account within 15 minutes on average.

The Kreditech website states that the company's algorithm processes up to 20,000 items of information per request; as well as the customary credit details, which are obtained from Schufa, it also uses GPS locational data, social graphs (Likes, friends, locations and posts), online shopping histories and device data (Schulz, Müller and Rosenbach, 2013).

Besides the information provided on the application form, the system also assesses the process of completing the form: did the applicant use an iPad, which would have been expensive, or another device? How long did it take to complete the form? The frequency of errors and the use of the delete key are also registered. Kreditech's founder, Sebastian Diemer, said in an interview with the daily newspaper *Die Welt*, "We observed that people who did not repay their loan had a very particular font on their computer"⁴⁴ (Seibel, 2015). Information about a person that is accessible on the Internet is also analysed. With the applicant's consent, social media such as Facebook and Twitter are also analysed, as are sites such as Amazon and eBay. How many of the applicant's Facebook friends have been to university? How creditworthy are they? Facebook profiles are also used to check whether the applicant's photograph and location match those on other networking sites such as Xing or LinkedIn.

Kreditech is only one example of a number of companies that use similar business models to grant or broker loans for their customers. Others include Zestfinance⁴⁵ and Kabbage,⁴⁶ which are based in the United States, Wonga⁴⁷ and Big Data Scoring⁴⁸, based in the United Kingdom, and Lendo⁴⁹, based in Singapore. In 2015, Facebook was granted a patent for a process which, among other things, is designed to enable lenders to assess the creditworthiness of Facebook members on the basis of their circle of friends (Lunt, 2014).

42 <https://www.kreditech.com/>, accessed on 18 June 2018.

43 In Germany, Kreditech withdrew its services after only three weeks when the Federal Financial Supervisory Authority (BaFin) announced an audit of its business model (Schulz, Müller & Rosenbach, 2013).

44 The font in question is one that is used by casino and poker programmes. These applicants, according to Kreditech, were therefore more likely to be online gamblers (Seibel, 2015).

45 <https://www.zestfinance.com/>, accessed on 18 June 2018.

46 <https://www.kabbage.com/>, accessed on 18 June 2018.

47 <https://www.wonga.com/>, accessed on 18 June 2018.

48 <https://bigdatascoring.com/>, accessed on 18 June 2018.

49 <https://www.lendo.com/>, accessed on 18 June 2018.

All of these business models are characterised by the use of complex big-data analyses and algorithmic decision-making processes, and they show what has now become technically possible. As described above, the data that are used go far beyond the traditional variables such as payment history and entries in debt registers.

Scoring in the health sector

In the health sector too, we are seeing the development of an increasing number of business models which, though not always based on big data, do at least use algorithmic scoring methods to quantify people's state of health or health-consciousness and/or to make predictions about their future health (Kolany-Raiser, 2016; Wiegard and Breitner, 2017; Budzinski and Schneider, 2017). Such products are of particular interest to health insurers if they serve to guide and monitor policyholders' behaviour and to influence the cost risk and make it calculable. Among the data sources are smart bands and other wearable devices that measure health parameters such as blood pressure and heart rate as well as tracking sporting activity and exercise. Even sleep patterns and diet can be recorded by means of apps and wearable devices. Our research, however, has shown that scoring projects in the field of health insurance, at least in the highly regulated German market, have hitherto been chiefly focused on customer acquisition and retention (see chapter C.III below).

One of the private service providers in the health service Dacadoo,⁵⁰ a Swiss start-up business that was created in 2010 and has its headquarters in Zurich. Dacadoo measures a person's current state of health and well-being by means of the personal health index – a value ranging from 1 for the poorest state of health to 1,000 for perfect health. The index goes up or down in real time, depending on body readings, emotional well-being and lifestyle (exercise and sport, stress and sleep). Activities are recorded in the Dacadoo app with the activity tracker. Other fitness-tracking devices and apps are also supported. In addition, the Dacadoo app can be connected with digital scales, blood-pressure gauges and wrist-

worn heart-rate monitors. Lastly, data can be entered manually through the app or through the Web interface. Dacadoo also uses information on past health problems, including details of family medical history, on hospital stays and on lifestyle habits.

All of this information is provided by the Dacadoo company. There has not yet been a clinical study, however, to demonstrate that using the Dacadoo platform is actually effective in modifying people's behaviour and improving their health.

One illustrative example of a sophisticated system of health scoring by a health insurer is the programme operated by the insurance company Aviva in the United Kingdom. Anyone who takes out private insurance with Aviva can participate free of charge in a programme called MyHealthCounts.⁵¹ A person's current state of health is assessed by means of a rating known as a 'Q Score'. With the aid of tracking devices, policyholders' weight, blood pressure, fitness levels, smoking habits, cholesterol levels and more can be monitored. After registering, the policyholder also completes a questionnaire. This contains questions about factors like diet and exercise and the medical history of the policyholder and his or her family. On this basis a personal score is calculated. A person's health score is calculated as his or her position in an imaginary prioritised queue of 100 people of the same age, sex and race and sex at a doctor's surgery, numbered from the least fit to the fittest. The programme gives participants the opportunity to obtain up to 15% off their next renewal premium.

Aviva's partner for the MyHealthCounts programme is the British group Roadtohealth, which developed the health-scoring app Quealth.⁵² The app calculates each individual's risk of contracting various illnesses in the future. To this end, the user of the app answers questionnaires on his or her medical history, biometrics and lifestyle. Algorithms calculate the Quealth Score, expressed as a figure between 1 and 100. As with MyHealthCounts and Dacadoo, the precise algorithm is not known. The risk is currently assessed for five non-transmissible

⁵⁰ <https://www.dacadoo.com/>; accessed on 18 June 2018.

⁵¹ <https://www.myhealthcounts.co.uk/>; accessed on 18 June 2018

⁵² <https://www.quealth.co/>; accessed on 18 June 2018.

diseases – diabetes, cardiovascular diseases, cancer, dementia and chronic obstructive pulmonary disease. The company claims on its website that the Quealth algorithms are evidence-based with proven high predictive accuracy. However, no independent clinical study is known to have confirmed these claims.

As well as for risk assessment, the company says that it uses the collected information to provide customised health information, tips and health coaching.

Another example of the way in which machine-learning algorithms that are constantly trained with new data can be used in the health sector comes from the United States. SelfieQuote.com is a pilot project launched by life-assurance company Legal & General America.⁵³ Customers can obtain a quote for their life insurance by uploading a selfie of their face. The technology analyses the photo on the basis of a hundred attributes of the person's face and extrapolates proxy information on body mass index, gender and biological age. Another feature of the technology is the possibility of estimating whether the person is a smoker. According to Legal & General, the data obtained in this way are used to verify the statements made by the customer, and then a personalised insurance package is put together. The precise process is known to the company alone. SelfieQuote probably serves primarily as a marketing instrument at the present time, designed to recruit young, tech-savvy customers (Moorcraft, 2018).

The Legal & General project was created in partnership with technology firm Lapetus Solutions.⁵⁴ With a combination of facial analysis, biodemographic information and surveys, the project team claim to be able to calculate people's health status and life expectation or death risk. At the present time, they are examining the possibility of using technology to identify the first signs of illnesses such as diabetes, heart defects and dementia.

Another firm based in the United States is the start-up business Aspire Health, which is cofinanced by Alphabet, the parent company of Google.⁵⁵ Its main area of activity is the provision of services in the field of palliative care. Aspire Health has developed an algorithm designed to calculate how long a seriously ill person will survive. Health data on patients are collected, and their clinical symptoms are checked against the outcomes of frequently used therapies (Welchering, 2017). The aim is to determine which further treatment seems to be advisable. The algorithm delivers a recommendation on how the patient should be treated, and the physician makes the decision. The process is based on the assumption that expensive medical treatment is inadvisable if the patient's chances of survival are slim. Advocates of the approach believe that it can increase the profitability of healthcare provision.

Critical consideration, particularly as regards the predictive power and effectiveness of the various scoring models, is imperative, and not only in the cases described above, for as a rule there is a lack of independent scientific evidence that the scores actually do what they promise. Moreover, the consideration of profitability as a decision-making criterion in the realm of health care requires critical reappraisal. The cited examples show, however, that the quantification of health with the aid of algorithms is a versatile tool and is already being practised. It remains to be seen how the trend develops and whether similar business models can and will become established in Germany too.

53 <https://term.lgamerica.com/selfie-quote/#/>; accessed on 18 June 2018.

54 <https://www.lapetussolutions.com/>; accessed on 18 June 2018.

55 <http://aspirehealthcare.com/>; accessed on 18 June 2018.

2. Data accumulation and data trading

The foundations of any system of consumer scoring are consumer data, sometimes in considerable volumes (Dixon and Gellmann, 2014). There are scarcely any areas left in consumers' lives in which they do not leave a digital footprint. As a result of the digital revolution, it has never been so easy to collect personal data. Similarly, there has been a growing economic interest in making profitable use of these data, for example in the realms of advertising, e-commerce, market research and politics. The aim, in most cases, is to prevent fraud, to identify people or to make marketing campaigns and customer communication more efficient by pinpointing and addressing specific target groups (OECD, 2013).

Data that cannot be collected by the company that intends to use them are often bought in from data traders. The business model of such enterprises consists in collecting, processing and selling data. In 2014, the trade in addresses and other personal data was worth some 610 million euros (Goldmedia, 2017).

The personal data that are collected about consumers can be classed in various categories (OECD, 2013): demographic data (e.g. date of birth, gender, civil status, level of educational attainment and income), user-generated material (e.g. blogs, comments, reviews, photos and videos), Internet browsing history (e.g. search-engine queries and online shopping history), data relating to a person's social environment (e.g. contacts and friends in social networks), location data (e.g. address, GPS data and IP address) and official personal communication data (e.g. passport number, account numbers and police records).

As a rule, data traders do not obtain data directly from consumers but through third parties, and often without consumers being any the wiser. Data traders use numerous sources to build up their databases, for example

from online traders and mail-order firms, through customer loyalty cards or cashback cards, through competition entries, from publicly accessible statistics and directors or from data collectors such as Facebook, Amazon, Twitter or other website operators (Goldmedia, 2017).

Data collectors can obtain personal data on the Internet in many different ways (Palmetshofer, Semsrott and Alberts, 2016): on the one hand, there is a direct route in cases where data are voluntarily transferred, e.g. when users register with an online service or post their profile on a social network site. Then there are data which are gathered when a person is observed and so discloses information indirectly, e.g. through browsing history or GPS locations. Lastly, data such as online profiles, likes or reviews and even seemingly non-personal data are analysed and evaluated for the purpose of extrapolating personal data, such as an individual's age or sex.

It is virtually impossible for consumers to trace where their data goes. Most data traders receive some of their material from other data traders and sell it in turn to more data traders. A study conducted by the Federal Trade Commission (2014) found that seven of the nine US-based data traders under examination made data available to each other.⁵⁶ This means that consumers have few opportunities to object to any previous use of their data, to correct false data and to contest their classification in particular categories, because they do not know who holds their data and which of their data are already available globally for sale.

A key concept in this discussion is that of informed consent, which is the legal basis for the collection of many of the items of data referred to above. A consumer freely agrees, for example, to have his or her usage history recorded, which would otherwise be prohibited. Consumers normally express their consent by agreeing to a privacy statement or to a set of standard business terms. There are, however, some problems associated with informed consent (Hofmann and Bergemann, 2017).⁵⁷ On the one hand, studies show that most consumers do

⁵⁶ The nine companies examined in the study were Acxiom, Corelogic, Datalogix, eBureau, ID Analytics, Intelius, Peek You, Rapleaf and Recorded Future (Federal Trade Commission, 2014).

⁵⁷ The challenges described here illustrate clearly that a consumer's informed consent alone cannot be a key instrument of data protection but must be supplemented by other mechanisms. Accordingly, the guarantees provided by the principle of purpose limitation (Article 5(1)(b) GDPR) and the prohibition of conditionality (Article 7(4) GDPR) are crucial; on this point, see SVRV 2016.

not read the text of the privacy statement or the standard business terms. Ticking the appropriate box before downloading an app or joining a social network has become a habitual everyday action. One contributory factor is that it is mostly impossible for users to access the desired service without giving their consent. Users therefore see no real alternative that would not mean severely restricting their daily activities or opting out of social participation. Another criticism of informed consent is that it may be doubted whether consumers are fully informed at the time when they give their consent. Privacy statements and standard business terms often run to several pages and are worded in a way that is incomprehensible to non-experts. Most users, moreover, are entirely unable to assess the potential implications of consenting to the storage and processing of their data. This is why the SVRV, in its report *Digital Sovereignty* (SVRV, 2017a), advocated a rule requiring businesses to inform consumers of their standard business terms and privacy policy in an easily comprehensible one-page statement not exceeding 500 words. Some consumer bodies have also proposed the introduction of preformulated privacy provisions (Pollmann and Kipker, 2016).

Even though individual sources may only make a few items of data available about a consumer, data traders can accumulate many different items. In this way it becomes possible to obtain a detailed picture of an individual's lifestyle and behaviour. What is more, data traders not only aggregate raw data, such as a person's name, age, payment transactions and search history but also data elements derived from them, in other words conclusions regarding, for example, product preferences, purchasing power or fitness level but also extremely delicate categorisation into groups based on ethnic background, income brackets or health status (on the associated dangers of indirect discrimination, see section B. II.2 above). A study in the United States, for example, has demonstrated that an analysis of someone's Facebook Likes can provide a fairly reliable guide to characteristics such as sexual orientation, ethnicity, religion, political attitudes, intelligence, narcotics consumption, age, relationships, gender and more (Kosinski, Stillwell and Graepel, 2013). Various personal characteristics can be

concluded, for example, from an analysis of mobile-telephone data (Chittaranjan, Blom and Gatica-Perez, 2011; De Montjoye, Quoidbach, Robic and Pentland, 2013) or search-engine use and browsing behaviour. (Kosinski, Stillwell, Kohli, Bachrach and Graepel, 2012).

Available information is also used to sort consumers into groups and categories. Assignment to a group does not normally depend on one single characteristic but on a combination of various items of information. The data trader Acxiom (see below), for example, uses categories such as "middle class", "active urban" and "status-driven working class" (Junge, 2012).

As was shown at the start of this chapter with the aid of some examples, novel scoring models are making increasing use of big data and algorithm-based processes. Since the computation of a score does not require a causal relationship between two different variables but only a statistical correlation, statistical methods are used to analyse large volumes of consumer data and to recognise patterns and connections with a view to making predictions or estimates of future consumer behaviour (Christl, 2014). There is therefore great interest in accumulating the largest possible volumes of data. Jeanette Hofmann fears that this might lead to a data-based creation of monopolies. "It is actually the case", she writes, "that the generation and analysis of data is in the hands of very few organisations. And of course it is also the case that possessing more data enables a company to construct better algorithms, which means that there is, so to speak, an inherent propensity for monopolisation in this market." (Hofmann, in Bilger, Löwel and Tomaszewski, 2018; see also, for example, Rubinfeld and Gal, 2017).

Examples of data traders internationally and in Germany

The Acxiom Corporation⁵⁸ is an enterprise from the United States; it provides marketing services to corporate clients, including the provision of consumer data. In the United States, Acxiom reportedly holds data on 250 million people. According to its website, Acxiom is active in more than 60 countries and promises its clients access to 2.5 billion

58 <http://www.acxiom.de/ueber-acxiom/>; accessed on 18 June 2018.

consumers⁵⁹. A 13-digit number is allocated to each person whose data it holds, and all stored information is linked to that number (McLaughlin, 2013): demographic data, household characteristics, financial situation, life events, interests, buying activities, social behaviour – the list of data elements is long.⁶⁰ Sorting consumers into particular target groups is as much part of Acxiom's services as the matching of data with specific e-mail addresses, postal addresses or telephone numbers. A particular speciality of the corporation is the acquisition of offline data from sources such as government authorities, with which the online data are amplified. Acxiom also offers its clients the storage and administration of complete customer databases. Its clientele includes credit-card providers, vehicle manufacturers, insurers, retail chains and many more.

Since 2004 Acxiom has been operating in Germany too, where it offers clients address data with numerous other items of additional information on the address in question and on the consumers who live there. The data sets can be tailored to specific consumption intentions, lifestyles and attitudes of consumers. Acxiom Deutschland has already collected data on 44 million Germans.⁶¹

Besides Acxiom Deutschland, the dominant companies in the data-trading business in Germany are ABIS GmbH, which is part of the Deutsche Post Address Group, AZ Direct GmbH, part of the Bertelsmann Printing Group, EOS Holding GmbH, a subsidiary of the Otto Group, and Schober Information Group Deutschland GmbH (Christl, 2014; Goldmedia, 2017). They mainly operate in the realm of address trading; that is to say they collect consumers' postal addresses, verify and process them and supplement them with additional items of information. They obtain the addresses through publicly accessible channels, such as telephone directories and registers of companies and associations. Data traders also buy data from other companies, such as mail-order firms and publishing houses.

German-based AZ Direct GmbH, for instance, administers 37 million private addresses and offers these to its clients for direct marketing activities.⁶² Through its AZ DIAS audience-targeting system it can provide profiling data on 40 million households, 70 million individuals and 20 million buildings, covering socio-demographic and psychographic attributes, consumption patterns, stages of life, location/geodata and much more. Prominent companies such as the Weltbild publishing group, the Gruner + Jahr publishing house and the Klinger industrial group have their address bases administered by AZ Direct. Visitors to the AZ Direct website⁶³ can look up descriptions of more than 2,500 lists that divide customers of client companies into categories such as 'environment-conscious', 'donation-prone retired academics' or 'socially minded multiple mail-order purchasers'.

59 <https://www.acxiom.com/>; accessed on 18 June 2018.

60 A selection can be found at <https://www.acxiom.com/what-we-do/infobase/>; accessed on 18 June 2018.

61 <https://www.linkedin.com/company/acxiom-deutschland-gmbh/>; accessed on 18 June 2018.

62 <http://www.az-direct.com/site/multichannel-marketing-produkte/direct-mail/>; accessed on 18 June 2018.

63 <http://www.az-direct.com/site/fileadmin/ikat/listinfos/index.html>; accessed on 18 June 2018.

3. Repersonalisation of anonymised data

Anonymised data, in other words data that are not re-traceable to a particular consumer, are not covered by the Data Protection Act and may therefore be collected and used in Germany without restriction and be freely bought and sold by data traders. The specific connection to a person – typically a name, date of birth and address – is therefore removed from the data sets. Another option is to pseudonymise the data, which means that attributes such as the name or address are replaced with a pseudonym or other identifier.⁶⁴

It would be wrong, however, to underestimate the danger of ‘de-anonymisation’ or ‘repersonalisation’ of such anonymised or pseudonymised data. It normally takes only a few items of related data to make individuals identifiable again. Sweeney (2000), for example, showed that zip code, gender and date of birth are sufficient to identify unmistakably 87% of the US population. County, gender and date of birth are already enough to identify 18% of US citizens.

There is ample evidence of the ways in which anonymised data sets can be repersonalised. Researchers at the University of Texas, for instance, using algorithms that they had developed themselves, managed to repersonalise parts of an anonymised Netflix data set with film reviews posted by 500,000 users. By reconciling the data with non-anonymised film reviews on the Internet Movie Database website, the researchers succeeded in identifying individual users. It also proved possible to uncover their political preferences and other sensitive information (Narayanan and Shmatikov, 2008). In 2009, the same authors demonstrated how active users of both Twitter and the image-hosting service Flickr could be reidentified from an anonymised Twitter data set with an error rate of only 12% (Narayanan and Shmatikov, 2009). Other studies have shown that anonymised mobility data obtained from GPS sensors in smartphones, computers and vehicles can be repersonalised by supplementing

the data sets with additional information, for example from social networks (see, for instance, De Montjoye, Hidalgo, Verleysen and Blondel, 2013; Ji, Li, Srivatsa, He and Beyah, 2016; Srivatsa and Hicks, 2012). In Germany, reporters from the regional broadcaster Norddeutscher Rundfunk (NDR) succeeded in repersonalising a dataset they had obtained from a data trader which contained some ten billion IP addresses, retrieved from about three million German Internet users (Norddeutscher Rundfunk, 2016). Detailed browsing histories proved to be assignable to specific persons. The data set, in fact, was apparently obtained by unlawful means: according to NDR, a Web of Trust browser extension from WOT Services had recorded the websites visited by users without obtaining the users’ consent and had then stored the data on servers outside Germany.

It may therefore be argued that, in the age of big data, every item of data is potentially personal because of the countless possibilities of linkage with other personal data (Boehme-Neßler, 2016). The collation and reconciliation of several data sets containing personal information such as shopping records, browsing histories, search histories, etc., from various sources make it possible to trace these back to specific consumers. Big data and powerful algorithms allow this to be done without any great effort. In this way, data acquired from data traders can also be used for consumer scoring, since those data are also compiled on individual persons as a rule.

To sum up, it may be said that the accumulation and trading of data play a major role in the age of big data and that it is possible with the aid of algorithms to identify specific consumers, even in large anonymised data sets. But what does this mean in relation to consumer scoring?

⁶⁴ The question whether pseudonymised data are not personal is a controversial one. Recital 26 of the GDPR states that “Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information, should be considered to be information on an identifiable natural person.”

4. Aggregation of data into a super score

As data trading grows in significance, so does the potential for data from the most diverse areas of people's lives to be brought together in a single database and a single company. Potentially, then, data from various supposedly unconnected areas of activity could be matched with particular consumers and then used as a basis for scoring them. Such a super score would mean that an individual person's behaviour in a particular context could have far-reaching implications for every area of that person's life (see section B.II.4 above).

Evidence of data and scores being diverted to other purposes has surfaced in the United States, for example. Credit information, such as the FICO score, is used by providers of motor and household insurance to calculate their premiums (Consumer Reports, 2015; Dixon and Gellmann, 2014). What interests these insurers is not how likely a customer is to repay a loan but rather the degree of probability that he will be prepared to pay higher insurance premiums or that he will make a claim, and the data are fed into these assessments (O'Neil, 2016).

In a survey conducted by the Society of Human Resource Management, almost half (47%) of the 430 respondent employers stated that they had a credit check conducted before they hired a new employee (Society for Human Resource Management, 2012). In this case too, the aim was not actually to find out about an applicant's creditworthiness but to infer attributes such as trustworthiness and reliability (O'Neil, 2016).

The diversion of data to other purposes is encouraged by data trading. Credit scoring in particular is already closely associated with the data-trading business – often in an inscrutable labyrinth of corporate conglomerates and subsidiary companies. Two components of the Bertelsmann media group, for instance, are the data-trading firm AZ Direct GmbH and the credit reference agency Arvato Infoscore. Creditreform Boniversum, as well as performing its own function as a credit reference agency, also administers, through its subsidiary Microm,⁶⁵ a consumer database registering socio-demographic, socio-economic and psychographic attributes. So there is undoubtedly scope for the use of more unconventional attributes, such as online behaviour, for credit scoring as well as for other forms of consumer scoring.

It should also be mentioned at this point that such scope also exists regardless of data traders, for example in major insurance companies that offer various types of insurance. With the consent of policyholders, data sets can be combined in these cases and analysed together, including behavioural data. The insurer Generali Versicherung AG, for instance, currently offers smart insurance and telematics-based options in four types of insurance: life insurance and occupational disability insurance (*Generali Vitality*), motor insurance (*Generali Mobility*) and household contents insurance (*Generali Domocity*). The *Vitality* programme is to be introduced in the near future in the realm of private health insurance. Even though it must be stressed that Generali has not given any indication at all of plans to combine the data sets, interesting questions certainly do arise in the light of the danger posed by super scores (see also the market study in Part C below): could driving data not be relevant in the calculation of health insurance premiums? After all, accidents involving personal injury may increase the cost of health care. And, conversely, does not a driver's high risk of heart attacks increase the probability of an accident?

⁶⁵ <https://www.microm.de/>; accessed on 18 June 2018.

The scenario of a super score, in which a person's behaviour in various areas of his or her life is analysed and used to calculate a score, seems entirely plausible in the depicted context of innovative business models, big data, data trading and de-anonymisation. While developments like the social credit system in China are not to be expected from the governmental side in Germany, they do raise the question whether consumers need to be better protected against similar developments in the business world. If large volumes of data on consumer behaviour are easily obtainable and analytical tools can be profitably deployed, data trading and consumer scoring will continue to proliferate. The fact that companies have a great interest in such developments is illustrated, for example, by a remark made by Douglas Merrill, founder of the credit reference agency Zest Finance; his analysis of the current state of play can also be taken as a warning: "We feel like all data is credit data, we just don't know how to use it yet" (Hardy, 2012).



Market survey: credit reference agencies, motor insurance telematics and health insurance policies



I. Introduction and key issues

In order to analyse the current supply of consumer scoring services in the German market in the three areas under examination in this report, namely credit scoring, telematics-based motor premiums and health scoring, the SVRV conducted a market study in the spring of 2018, surveying agencies and insurers in the relevant market segments.

The method of a standardised written questionnaire was chosen to obtain responses directly from insurers and other businesses and possibly learning more than research into company websites, etc., would reveal. Another objective was to supplement the examination of the consumers' perspective in Part D with a portrayal of the perspective and interests of compilers and users of scores.

One of the aims was to gauge the current prevalence of scoring systems and behaviour-based business models in the three areas under examination. The three key questions were:

- Which products are available and which are in the pipeline?
- Why are people scored? What objectives are being pursued?
- What data are used?
- Which quality criteria do scoring systems meet?
- How transparent are companies about their scoring systems?

The other aim was to give respondents the opportunity to communicate their experiences, opinions and plans relating to scoring and behavioural tariffs. We were interested, for instance, in learning what they saw as the pros and cons of scoring and how they saw the future of such systems in their respective sectors. With a view to obtaining the frankest possible responses, we assured the insurers and other business representatives that their responses would be analysed anonymously.

In the analysis of the findings, our intention was to study consumer problems in conjunction with corporate scoring models, with due regard to the problem areas identified in Part B above, in order to build an empirical basis for the recommendations for action that we make in this report.

II. Survey design

The market study conducted by the administrative office of the SVRV encompassed firstly credit reference agencies, secondly motor insurers and thirdly both statutory and private health insurers. Market research was conducted into these three sectors with a view to identifying relevant businesses for the survey, and a market profile was produced to serve as the basis for the survey.

Although specific consumer problems could be expected to emerge in each of the individual market segments, provision also had to be made for an overarching discussion. For this reason, care was taken when conducting the survey to maintain a sufficient degree of standardisation to ensure that most of the questions were identically or similarly worded for all respondents in the three areas (see Annex II).

At the same time, specific features of the three segments also had to be taken into account. For example, the survey dealt with potential future trends and influencing factors that varied from one segment to another. For the credit reference agencies, for example, there was a question about the possible use of data from social media, motor insurers were asked for their views on the European eCall Regulation, and health insurers were asked about electronic patient files.

The consideration of specific conditions in the survey was most evident in the domain of health insurance, because neither statutory nor private insurers in Germany offer policies with lifestyle-based premiums at the present time. Policyholders may, however, take part in bonus programmes in which healthy activities and participation in preventive measures are 'scored' in a sense,⁶⁶ inasmuch as they earn bonus points which policyholders can redeem for monetary or other rewards. In the cases of credit reference agencies and telematics-based motor insurance tariffs, however, the degree of automation is considerably more advanced. The questionnaire for insurers focused far more on establishing where ventures into health scoring were already in evidence and what the respondents' wishes, plans and views were with regard to health scoring.

Health insurers, unlike credit reference agencies and motor insurers, were not confronted with the specific term 'scoring', since the concept of 'lifestyle-based tariffs' seemed to be commoner in this sector and therefore more expedient. For this reason, most of the questions relating to score calculation, modelling, statistical quality criteria, etc., were omitted from the health insurers' questionnaire.

On the whole, then, a comparative analysis of the survey findings was possible, while specific conditions and potential trends in a particular sector could also be addressed. In the analysis, scoring-related consumer problems affecting all three market segments could be highlighted, and potential best practices could be elicited.

The questionnaire was not pretested but was developed on the basis of background discussions with representatives of agencies and insurers and with scoring experts.

A word of caution: the findings of this market study are based solely on responses from representatives of agencies and insurers, and it was not always possible to validate these, for example by consulting publicly available sources. The primary aim of the study was therefore to establish the nature and extent of the information that agencies and insurers make available about their scoring systems and also to find out what views, ambitions and wishes they would express with regard to scoring when assured of anonymity.

For this reason, most of the questions were open-ended, that is to say there were only a few questions in which respondents were able to select a preformulated reply. For the analysis of the open-ended questions a quantitative content analysis was conducted, and a classification system with summative categories and associated codes was developed. All responses were coded by three estimators acting independently of each other.⁶⁷

⁶⁶ Some bonus programmes even offer the option of using data from fitness apps to earn points as well as enabling participants to check their current points total and administer their points account digitally.

⁶⁷ Where responses were coded differently, the category on which two estimators agreed was selected.

In the areas of credit scoring and telematics-based motor insurance tariffs, the response rates both exceeded 50% (60% for the credit reference agencies and 63% for the motor insurers), while the response rate for health insurers came to 41%. The findings cannot be considered representative of the respective sectors in their entirety, and the responses reported here provide limited scope for generalisations.

1. Overview of providers

Credit reference agencies

To analyse the German market in credit scores, we surveyed various credit reference agencies. These are private-sector companies which communicate to business partners economically relevant data and credit ratings pertaining to individuals. Questionnaires were sent to five agencies, all of which are members of the association of credit reference agencies known as *Die Wirtschaftsauskunfteien e. V.* (formerly called *Verband der Handelsauskunfteien e. V.*) and provide credit information on private individuals. Annex I.1 contains a list of the participating firms.

Motor insurers

Another questionnaire was sent to motor insurers that offer telematics-based tariffs. There are no providers of motor insurance whose product portfolio consists entirely of telematics-based policies; on the contrary, telematics-based premiums are add-ons that are offered for existing motor insurance policies

A total of 15 telematics-based tariffs or add-ons were identified in the realm of motor insurance. Of the providers we contacted, five took part in the survey; some of these companies also responded on behalf of subsidiaries. As part of a pilot project implemented between 2013 and 2015, Sparkassen Direktversicherung had tested a telematics-based tariff and also took part in the survey. In total, responses relating to ten telematics-based tariffs were received. These tariffs are listed in Annex I.2.

Statutory and private health insurers

This market study also involved a survey of all health insurers operating in Germany. Although health scoring as such is not practised here, the aim was to shed light on the various bonus programmes so as to establish in what form and to what extent policyholders' healthy lifestyles are being registered and rewarded. In addition, insurers were questioned about their attitudes to lifestyle-based tariffs and on their plans for the near future with regard to such schemes.

At the present time, a total of 110 providers of statutory health insurance⁶⁸ and 43 private⁶⁹ health insurers are operating in the German market, of which a total of 62 (47 statutory and 15 private) took part in the survey⁷⁰ (see Annex I.3).

68 Information from the National Association of Statutory Health Insurance Funds at <https://www.gkv-spitzenverband.de/>; accessed on 10 July 2018.

69 Information from the Association of Private Health Insurance Companies at <https://www.pkv.de/>; accessed on 10 July 2018.

70 The participating Local Health Insurance Funds (AOKs) answered some of the questions identically, namely those relating to views on lifestyle-based tariffs and future prospects. When the responses were analysed, however, each response from each of the AOKs was assessed once, which meant that the responses in question were each assessed eleven times.

2. The questionnaires

A separate questionnaire was compiled for each market segment (see Annex II). The questionnaires were sent by post and e-mail to the providers referred to above at the end of March 2018. Assurances were given that responses being analysed would not be directly identifiable with the name of the responding company.

In particular, questions were framed on the following subjects:

1. business models,
2. data collection and calculation of scores,
3. oversight and information rights,
4. views on the subject of scoring and future prospects

First of all, the insurers and credit reference agencies were asked a few questions about their business model. Among the items of information we were interested in learning from the credit reference agencies were the conditions on which credit scores are disclosed. As regards the motor insurers' telematics-based tariffs and the health insurers' bonus programmes, the questions were designed to elicit information such as whether these schemes were focused on a particular target group and what attributes characterised the customers who took part in them. There were also questions on the benefits for participants and the impact on their driving or on the healthiness of their lifestyle.

There was also a set of questions on data collection and the calculation of scores for the credit reference agencies and the motor insurers. Which sources or technology are used by companies to collect data (e.g. plug-in telematics modems or smartphone apps) was of interest to us, as was the question of who calculates the scores. Then came questions about the input variables and about the estimation models that are used and their quality.

Questions on supervision and quality assurance covered measures taken to guarantee that the input data were up to date as well as the existence of accurate and procedural checks by supervisory authorities.

Following questions on customers' information access rights and their right to have errors corrected, the insurers and credit reference agencies were asked for their appraisal of the current state of their respective markets, of scoring systems and of possible future developments. They were asked which items of additional data companies would like to collect and evaluate and whether they saw a need for adaptation of laws and regulations. They were also asked to discuss the advantages and disadvantages of scoring. The question whether there are already specific plans to introduce behavioural tariffs and/or whether these have been judged beneficial is particularly relevant to the realm of health insurance, since health insurers do not yet offer such tariffs.

III. Discussion of findings and highlighted consumer problems

A comprehensive description of all of the findings of the market study is set out in Annex III. In addition, Annex IV contains a tabular presentation of the findings. This chapter discusses the key findings and the consumer problems they highlight as these relate to the areas for action examined in Part B above.

1. Diffusion of scoring in the market segments under examination

While credit scoring by credit reference agencies is an established line of business, the market study shows that scoring has made far fewer inroads into motor and health insurance than might be inferred from the public debate.

Of the 90 or so firms that currently market motor insurance policies in Germany,⁷¹ our research showed that only 16 offer telematics-based tariffs in which the cost of premiums is determined by a score reflecting the policyholder's driving habits. These include major insurance groups with large numbers of policyholders, yet the survey revealed that the percentage who have signed up to a telematics-based tariff is still very low – less than one per cent for some companies. Six of the ten providers taking part in this survey target only young drivers and learner drivers with their telematics-based tariffs, because premiums traditionally tend to be high for this group and a potential cost reduction is particularly appealing to them.



“The objective reason for this product format is that young adults below the age of 25 have more frequent and serious accidents than drivers in older age brackets.”

Must telematics-based motor insurance be seen as a niche product, or will it become more widespread? Both assessments feature in the companies' responses. Some providers take the view that the use of telematics will radically transform business models in the industry and that the trend towards behavioural tariffs is irreversible.



“We believe that today's customers expect more from their insurer than claims settlement.”

One argument that is often advanced against the use of telematics from a corporate viewpoint focuses on the high costs arising from telematics-based tariffs for insurers. Especially in the case of business models involving built-in black boxes, the cost and effort of installing the technology and hardware, combined with the discounts for good driving, seem to make such schemes unprofitable, as responses to the survey suggest. This may also be the reason that some insurers have opted against installed hardware altogether and record data exclusively through a smartphone app. Some providers see a solu-

⁷¹ Information from the German Insurance Association (GDV) at <https://www.gdv.de/de/zahlen-und-fakten/versicherungsbereiche/ueberblick-4660>; accessed on 10 July 2018.

tion to the cost problem in the use of eCall technology,⁷² which, from this year, is being installed in every new vehicle. The respondents' frequently expressed fears of a data monopoly for manufacturers must be seen in this context.



“The pricing of motor insurance products is tailored very well to the risk profile, thanks to the pricing variables that are already in use today. It therefore remains to be seen whether telematics-based tariffs will actually replace the existing pricing variables or can at least supplement them.”

Behaviour-based health insurance tariffs, in which premiums are calculated on the basis of the policyholder's continuously monitored, i.e. scored, health-related behaviour, do not exist in Germany. It is, however, possible to obtain rewards from almost every statutory health insurance fund on submission of evidence of participation in healthy activities or in health-promoting measures. Private health insurers seem not to have established comparable bonus programmes.⁷³ It appears, moreover, that only a few policyholders avail themselves of bonus programmes. Indeed, the responses to the survey show that most health insurers have fewer than 10% of their policyholders on bonus programmes.

Nevertheless, according to the market study, the first signs of health scoring are emerging in these bonus programmes, especially those in which participants earn points, i.e. numerical credits, for their lifestyle choices.

Unlike credit scoring and telematics-based schemes, bonus programmes have generally been measuring the health-related behaviour of policyholders by non-digital means such as voucher booklets and participation certificates and by tallying the points that have been earned. Scarcely any use is made as yet of tracking by means of wearable devices or apps, even though an interest in such a development emerged clearly in various parts of the survey. No use is made, however, of complex statistical processes or even algorithms to establish a statistical correlation between lifestyle choices and individuals' health, which is why there is an urgent need to ask some fundamental questions about the fairness of these bonus programmes.



“Digitisation, coupled with the trend towards self-improvement, will impact on the health sector too. Scepticism towards data collection and transfer in the case of digital applications and therapies will decrease. When it becomes an everyday occurrence to disclose their health data, customers will also expect to receive individualised prevention offers and healthcare services. In future there must therefore be more scope to devise services and tariffs of this kind.”

⁷² Since 31 March 2018, manufacturers have been required to install eCall (short for 'emergency call'), the automatic in-vehicle emergency calling system prescribed by the European Union, in all new models of private cars and light utility vehicles. Devices installed in the vehicle automatically report a road accident to the standard European emergency number 112, thereby reducing the number of road deaths through the speedier initiation of rescue measures. The introduction of eCall requires, among other things, the installation in each vehicle of an airbag sensor, which signals whether an accident with a risk of injury has occurred, a satellite receiver, enabling the vehicle to report its location at the time of the accident by GPS and/or Galileo (this is often used in conjunction with the navigation system), a mobile-phone antenna to send the call to the emergency call centre and an eCall control unit (sometimes integrated into the in-vehicle infotainment system), which collects the data required for the e-call and establishes phone contact with the emergency call centre.

⁷³ The insurer Generali, however, plans to offer its app-based prevention programme *Vitality* as part of its health insurance package from late 2018 or early 2019 (<https://www.generali.de/ueber-generali/presse-medien/pressemitteilungen/generali-vitality-wird-in-deutschland-weiter-ausgebaut-25562/>; accessed on 1 October 2018).

Although health scoring in the narrower sense is not yet practised in the German insurance market, the insurers' responses indicate that it is becoming an increasingly relevant topic. Some of the respondent insurers expressed openness to the idea of behavioural tariffs. Five out of 62 insurers, i.e. 8%, stated that they saw no disadvantages in the introduction of behavioural tariffs.

Many insurers would also like greater freedom with regard to the use of fitness trackers and rewards for community-minded behaviour, even though it does not benefit the individual's health.

More than 40% of the respondent health insurers could imagine using data from their policyholders' electronic patient files for bonus programmes or behavioural tariffs. If bonus programmes are also considered in the light of the schemes offered by commercial providers such as Dacadoo and private insurers such as Generali, the potential for the development of health scoring becomes clearly recognisable, as does its growing relevance to consumers.



“The German health market will not be able to ignore it either – everything just takes longer in Germany.”

2. Transparency

Consumer-friendly scoring implies transparency regarding the attributes that are used to calculate a consumer's score. The survey revealed significant differences between the three market segments under examination as regards the transparency of their scoring variables.

In the case of credit scoring, it emerged clearly that it is difficult for consumers to fathom which data are collected for the purpose of credit scoring and precisely which of these data are used as scoring inputs (for the consumers' view, see Part D of this report). It may be said that there are differences in the data inputs used by the various credit reference agencies in their scoring models. In the market survey, these differences in input data emerged, for example, in the agencies' responses on the number of variables they use. One agency gave eight as the minimum number of attributes it used, while another quoted the figure 25. The number of scoring attributes also varies between market segments.

The differences between credit reference agencies in the data they use for credit scoring result to some extent from the variations between the data sources that are available to them. While all of the agencies have access to publicly accessible registers and directories and official notices, the same does not apply to information from individual contractual partners and cooperating bodies such as financial institutions, mail-order firms and other traders, telecommunications companies or energy suppliers. No credit reference agency publishes an exhaustive list of the variables it uses in the computation of scores.

The survey findings indicate that customers on telematics-based tariffs are generally well informed of the criteria that affect their score. Between three and nine characteristics, such as braking and acceleration habits and speed, normally form the basis of the scoring models. According to the information provided by the surveyed insurers, customers are informed about the scoring structure in the policy documentation (standard business terms, conditions of motor insurance, terms of use, privacy statement, etc.) as well as on the insurers' websites and telematics apps.

According to the survey responses, most providers offer their pay-as-you-drive customers the prospect of assistance in improving their driving techniques and of a general improvement in road safety. The names given to their programmes, such as *BetterDrive* and *Secure Drive* also suggest that participation in a telematics-based scheme will improve their own driving and that the score reflects how safely they drive. Having presented their telematics-based policies in this way, insurers leave themselves open to criticism when they include factors like time of driving (e.g. day or night driving) or driving location (e.g. urban or rural driving) in the calculation of their scores, because these are often beyond the driver's control. Of the ten respondent providers, nine take account of the day or time, seven consider the type of road, five the location, two the population density and one the distance or duration of the journey. All of these variables are parameters that a driver can scarcely influence, and so it cannot be said that the score depends solely on driving style, such as how the driver brakes, accelerates, takes bends, etc. Differences between the advertising promise of becoming a better driver and the real purpose of the score, namely reducing the probability of an insured event, pose problems in terms of consumer interests (see section B.I.5 above).

As with the motor insurers' telematics-based tariffs, participants in bonus programmes offered by statutory health insurers are told which activities can earn them points and when they start to qualify for rewards, otherwise the programmes could hardly modify their behaviour. The survey showed that many bonus programmes are based on a lengthy catalogue of measures and activities. According to a report from the North Rhine-Westphalian Consumer Advice Centre (Verbraucherzentrale Nordrhein-Westfalen, 2015), bonus programmes are impenetrably complex in the eyes of policyholders and make it difficult for them to gauge "how many and which measures are likely to be eligible and how much they stand to gain in their own specific circumstances". As with the pay-as-you-drive policies, a problem with some bonus programmes is that, while participants are offered the prospect of health improvements, credits are also awarded for public-spirited actions such as donating blood and attending first-aid courses. It would be worth trying to obtain an unambiguous representation of what the bonus programme is supposed to achieve.



"The programme is intended to support responsible and safe driving by young motorists."

In terms of transparency, it also matters whether, how and how often consumers are informed of their current score. Similarly wide divergences are evident between the three examined market segments in this respect.

In the case of credit scoring, consumers can request a free copy of their personal credit record to find out what data of theirs a credit reference agency has stored and what their current score is. The point, here, however, is that the initiative must come from the consumer. The consumer must make an active effort to find out which body holds which data. Given the limited degree of public familiarity with the various credit reference agencies, this must be seen as a problem. A consumer survey conducted by the Independent Data Protection Centre for Schleswig-Holstein and the GP Research Group (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein und GP-Forschungsgruppe, 2014) revealed that Schufa was by far the best-known agency, with a 98% recognition rate among the respondents, followed by Creditreform Boniversum, with 33%. The other credit reference agencies were known by far fewer than 20% of the respondents. Many consumers, then, are not familiar with the role of credit reference agencies in economic life, and they do not know when an agency is scoring them or which agency it is, for instance when automatic credit checks are made in connection with online purchases

In the case of pay-as-you-drive motor insurance policies, consumers can, as a rule, use the insurer's app at any time to check their current score. After each journey or when sufficient journey data have been collected, the score is updated. Policyholders can, in theory, find out the effects of individual journeys on their score.

Participants in the bonus programmes of health insurers often discover their status in the form of a points total. There are variations, however, in the frequency with which the current total is communicated to them. With many policies, the insured person has the completion of activities and measures certified by non-digital

means such as a stamp or written entry in a voucher leaflet, bonus pass or the like or submits an attendance certificate to the insurer. It is equally common for the current points total not to be communicated until a reward is paid out. Seven insurers offer policyholders the option of checking their points total with an app or through an online gateway.

It seems consumer-friendlier to enable consumers who wish to do so to check their latest score or points total at any time. There is therefore evident scope for development, especially in some health policies.

3. Score calculation and statistical quality

Scoring systems are intended to predict, as accurately as possible, a particular mode of behaviour or particular consequences of a mode of behaviour.

In credit scoring, according to the responses of the three credit reference agencies participating in this survey, the target variable is the occurrence of a negative credit event within a certain period.

In the case of motor insurance policies, the aim is to predict the probability of a claim event or the frequency of future claim events.

The stated objective of health scoring, besides guiding consumers towards healthier lifestyles, is to predict the future health status of the scored person. Even in the bonus programmes these dimensions are discernible to some extent. To fulfil their statutory mission, health insurers are expected to promote healthy lifestyles. According to more than 70% of the respondent health insurance funds with bonus programmes, participants are led to expect this effect. Connected with the aim of promoting a healthy lifestyle is the perception that healthier lifestyles will result in a future reduction of healthcare

costs. People who now behave in a health-conscious way would have fewer or less serious bouts of illness in future and therefore generate less expenditure for their insurers. Accordingly, the general principle in bonus programmes is that the more listed activities in which a person engages, the higher his bonus will be. What this implies is that the more points a participant earns, the less his future health care will cost.

Algorithmic processes can be used to compute scores. The usual method for the calculation of credit scores is multiple logistic regression. One credit reference agency, when answering the question about its estimation method, also selected the options 'decision trees', 'ensemble methods' and 'neural networks/deep learning'.

Logistic regression also appears to be in frequent use for the calculation of telematics-based scores.

No complex algorithmic decision-making procedures are used in the health insurers' bonus programmes. Where there are points systems, a policyholder's bonus points are simply added on to his or her points total. Health scores, which involve considerably more sophisticated processes, similar to those used by credit reference agencies and providers of telematics-based policies, are only available so far from private insurers such as *Dacadoo* and *Generali Vitality*.

The information provided by the respondents from all three areas of activity under examination on the statistical quality of the scores they use were not very revealing. There are various quality criteria available for the calculation of scores. These describe how good the various models are at classifying consumers in terms of risk, whether of loan default, road accident, illness or other event.

Some information on the quality criteria used by the credit reference agencies was provided in the questionnaires; all three agencies use the Gini coefficient. One agency stated that the this coefficient reached a value of 0.60 to 0.75, while another referred to a value of 0.60 to 0.80.⁷⁴ These statistics, it must be said, tell us nothing at all about differences in predictive quality be-

⁷⁴ The Gini coefficient would have a value of 1.00 for a perfect predicting instrument. A coefficient of 0.00 means that a correct prediction would be pure chance (see chapter B.IV).

tween groups of consumers, in other words the extent to which fairness, in its various dimensions, is achieved (see chapter B.VI above).

The respondent motor insurers did not answer the question on the use of objectifiable quality criteria such as the Gini coefficient for the correct identification of ‘good’ and ‘bad’ drivers.

The quality of the scoring model depends on the variables that are used to calculate scores and on their relative weighting. Scorers should use only variables that have a statistically significant influence on the target variable. Factors with a higher causal influence than others should be more heavily weighted, and vice versa. The credit reference agencies and motor insurers were therefore asked to name the criteria they used to select and weight variables. We were also interested to know how health insurers chose the activities and measures for inclusion in their bonus programmes and how they decided how many points each activity or measure would earn.

In the case of credit reference agencies, it is difficult to obtain information on the weighting they apply to their chosen variables, because they invoke trade secrecy (see sections B.I.1 and B.I.3). The agencies did state, however, that statistical significance underlay both their selection and their weighting of predictive variables. If a variable will increase the predictive accuracy of the score, it is inserted into the model. Weighting is based on the extent of the statistical correlation between the predictive and the target variable. An interesting point in this context is that one of the respondent agencies listed additional criteria for the choice of variables, namely an assessment in the light of data-protection law, availability, cost and the stability and reliability of the variable. If, for example, there are privacy reservations about a variable, it is not used in the calculation of scores, even if it possesses a certain predictive force. No specific comments were made on data quality. Its very mention indicates that it has hitherto been underestimated in academic literature, which strongly emphasises the formal characteristics of scores (cf. Hand, 2005; Hand, 2006; Verbeke, Dejaeger, Martens, Hur and Baesens, 2012; Britz, 2008).



“There are attributes which are significant but which it is not so easy for human understanding to grasp intuitively at first. When it comes to such attributes, a decision will have to be taken whether to include them in a model or not.”

It is customary in the motor insurance industry for those insurers who offer telematics-based tariffs to analyse existing statistics on driver behaviour and accident causes from bodies such as the Federal Statistical Office or the Royal Society for the Prevention of Accidents in order to identify the variables that are suitable for use in predicting the probability of a claim event. According to the information provided by the insurance companies, the statistical significance of each variable is taken into account when a model is devised.

Statements made by some telematics providers that human intuition and rules of thumb developed through corporate experience play a part in the choice of variables were considered in chapter B.V above.

In the bonus programmes offered by health insurers, policyholders are led to expect that participation will impact on their health by preventing illness. If the quality criteria discussed here were applied, every activity earning a bonus should actually promote good health, and the number of points awarded for each activity should reflect its relative beneficial impact. If one measure does more to promote health than another, it should also earn more points.

According to the market survey, when selecting eligible measures for their bonus programme, many health insurance funds are guided by whether the measure promotes health. In so doing, they follow the legal requirements set out in section 65a of Book V of the German Social Code and the rules of the Federal Insurance Office and of the National Association of Statutory Health Insurance Funds⁷⁵. A few health insurance funds, however, also focus on other criteria, such as market trends and the appeal of the activities for policyholders.

⁷⁵ *Leitfaden Prävention*, published by the National Association of Statutory Health Insurance Funds, lays down “the substantive areas of action and quality criteria for the services of health insurance funds in primary prevention and occupational health promotion” (https://www.gkv-spitzenverband.de/krankenversicherung/praevention_selbsthilfe_beratung/praevention_und_bgf/leitfaden_praevention/leitfaden_praevention.jsp; accessed on 10 July 2018).

The Federal Insurance Office, in its special report on competition in statutory health insurance, comes to the conclusion that the health benefits of some bonus programmes are not quality-assured, “because a bonus is not only attached to certificated prevention services as defined in *Leitfaden Prävention*, the prevention guidelines drawn up by the National Association of Statutory Health Insurance Funds in accordance with section 20(2) of Book V of the German Social Code, but ‘comparable’ quality-assured services within the meaning of section 65a(1)(3) of Book V of the German Social Code, such as activities culminating in the award of a sports badge or gym membership, also suffice to earn a bonus (Bundesversicherungsamt, 2018).

It is particularly interesting to note that some health insurance funds allow their policyholders to collect bonus points by engaging in activities for the common good, which do not promise any impact on the individual’s health and therefore have no preventive element. Such programmes are nevertheless presented as promoting health. Another question that arises is whether credit for altruistic actions is covered by section 65a of Book V of the Social Code. Five of the surveyed insurers reward blood donations, first-aid courses, bone-marrow typing, organ-donor cards and living wills. The reason given for this approach was that insurers want to reward this altruism and the fact that it benefits other people’s health.



“Those who do not think only of themselves but are actively committed to other people’s health have earned their bonus points.”

The Federal Insurance Office, which exercises regulatory oversight of the federally regulated statutory health insurance funds, does not authorise the inclusion of activities for the common good because of the lack of benefit to the individual’s health (Bundesversicherungsamt, 2018), which is why only a few bonus programmes operated by insurance funds under the direct authority of a federal state (Land) can reward such activities. It also emerged in

the survey that some insurers harbour a desire to obtain wider latitude in this respect from the Federal Insurance Agency. This also applies to the use of fitness trackers. At the present time only three of the surveyed health insurance funds make provision for the collection of bonus points with data recorded on fitness trackers. The main data item logged by these devices is the number of steps taken, but heartbeat and calorie consumption are also recorded. From a consumer perspective, the question is how quality-assured such approaches are. Do the health insurance funds have evidence that the activities recorded with fitness trackers really do have a preventive effect? Or do the targeted effects actually relate to marketing and customer retention?

The weighting of activities in bonus programmes, it seems, is not necessarily determined by their health-promoting effects. Many policies involve no weighting, all measures earning the same number of points. The insurers’ most frequent response to the weighting question referred to time and effort and to simplicity. The aim was to make it as easy as possible for insured individuals to collect points in the bonus programme.

If there are errors in the baseline data, this may directly affect the reliability of the score and have far-reaching consequences for consumers. For this reason we wanted to know from the firms that engage in credit-scoring and those offering telematics-based insurance policies what means are available to consumers to have errors in their data corrected.

All of the credit reference agencies offer consumers the opportunity to have data corrected if the latter notice that the agency possesses incorrect or outdated information about them. They cannot exercise this right, however, unless they know about the incorrect data (see section B.I.3 above), which is why we refer again to the transparency problem relating to credit scoring (see above).

Best practice continues to be observed by those surveyed companies which, according to their own responses, engage in continuous data monitoring and adhere to retention time limits so as to preserve the quality of their database. One credit reference agency lists among its activities training courses at its clients’ premises, regular data analyses, automated cross-checks and

regular reconciliation and inventories of stored data. In addition, the agency states that its contractual partners are under an updating obligation, in other words they must report changes made to their databases.



”If consumers alert us to incorrect data, the data are immediately corrected.”

Besides the companies’ responses to the survey, it should be mentioned that one credit reference agency, Schufa, also permits consumers to have recourse to an ombudsman – an external, neutral mediator – in the event of problems, on condition that the consumer has tried unsuccessfully to resolve the matter with the agency’s customer service department. According to the Schufa Ombudsman’s latest annual activity report, 984 submissions were made by consumers in 2017, compared with 1,017 in 2016. Of these submissions, 366 were admissible, and in 42 cases were decided in favour of the consumer. A total of 618 submissions were found to be inadmissible, because the consumer had not met the requirement of prior contact with the Schufa service centre for private customers. In most cases, the aim of the submissions was to obtain the early deletion of one or more adverse entries (Schufa Holding AG, 2018b). The 2017 activity report of the Data Protection Commissioner for Hesse states that the complaints relating to Schufa concerned cases in which adverse credit information had been linked to the wrong person. This, according to the report, had been caused by manual processing (Der Hessische Datenschutzbeauftragte, 2017).



“It is therefore important, on the one hand, to do everything possible to recognise and preclude any such measurement errors and, on the other hand, to ensure that there are ‘tolerances’, so that one item of incorrect data cannot instantly destroy a motorist’s score. And finally, let it be perfectly clear that every driver will have some incorrect data in his records, which means that every motorist is affected by inaccuracies and that allowance has already been made for these when the algorithms were calibrated.”

When telematics customers discover errors in the recording of driving data, they should also have these checked by their insurer and, if necessary, corrected. Not all providers of telematics-based motor insurance allow their customers to do this, as the survey shows. The main reason why this poses problems is that companies are known to be prone to error to some extent. Cases are reported in the survey, for instance, in which incorrect driving data is collected because the maps are outdated or because of interference affecting the GPS signal or the transmission of data.

4. Behavioural effects

When scoring systems have a behavioural element, the question is whether the desired effects can really be achieved by this means. Does the incidence of accidents decrease when policyholders opt for telematics-based motor insurance tariffs? Does people's state of health improve as a result of participation in their insurer's bonus programmes? The insurers were asked whether they themselves had any knowledge of such effects and, if so, how they obtained it.

"Initial analyses support the hypothesis that customers who have chosen the telematics module drive more foresightedly than the average customer without the telematics module."

Motor insurers' normal practice, it seems, is to compare customers with and without the telematics add-on in terms of frequency of accidents or total cost of claims. Policy providers do not appear to form a 'genuine' control group to ensure comparability across the whole range of relevant attributes.

As one motor insurer explicitly admits, it cannot be ruled out that motorists who already drive carefully and safely will be more willing to opt for a telematics-based policy. If comparisons are made between claim rates for participants and those for non-participants, the results will therefore be skewed. The frequency of claims made by the observed individuals was low anyway and did not fall solely on account of their participation.

"Customers who know that they are careful and responsible drivers are also more willing to sign up for a telematics-based tariff."

A similar approach is also in evidence among the health insurance funds. Most insurers testify indirectly to the health effects of their programme by citing the drop in expenditure on healthcare services or the cost of health care for participants in the bonus programme compared with the healthcare costs of non-participants. Under section 65a of Book V of the German Social Code, expenditure on bonuses for particular measures must be funded in the medium term from savings and efficiency gains resulting from those measures. The criteria for the evaluation reports with which the health insurance funds meet their accountability requirement were communicated by the Federal Insurance Office in 2005 in a circular^{76,77} addressed to the federally regulated health insurance funds. The following are among the minimum requirements laid down by the Federal Insurance Office:

- A "non-randomised controlled study" is necessary. It must be a full study and not just a sample survey.
- Matching techniques must be used. Control groups must be sufficiently large and must match the main socio-demographic attributes of the participants. The specific prescribed matching criteria are age, sex, insurance status, region and costs.
- A 'before and after' comparison must be made.
- All costs (for the creation, implementation, documentation and evaluation of the bonus programmes) must be taken into account.

⁷⁶ https://www.bundesversicherungsamt.de/fileadmin/redaktion/Risikostrukturausgleich/Rundschreiben/Rundschreiben_bonus.pdf; accessed on 1 October 2018.

⁷⁷ The Federal Insurance Office acknowledges that it has to rely on baseline data from the health insurance funds when conducting cost-efficiency checks and that it can only check whether the evaluation reports are complete and plausible and the requirements set out in the circular have been fulfilled (Bundesversicherungsamt, 2018). According to the Office's activity report for 2016, only eight out of 66 federally regulated health insurance funds had been unable to furnish evidence of the past cost-efficiency of their bonus programmes (Bundesversicherungsamt, 2016).

These requirements are methodologically sound, especially as costs prior to the selection of a bonus tariff have to be considered. This means that the bonus-programme group is to be compared with groups which generated the same average costs at a time when neither group was participating in the programme. Whether the health insurance funds fulfil these requirements, particularly as regards the proper conduct of control-group comparisons, is difficult to establish. In the survey the health insurance funds did not provide any details of their evaluation procedures. The Federal Insurance Office acknowledges that it has to rely on baseline data from the health insurance funds when checking cost-efficiency and that it can only check whether the evaluation reports are complete and plausible and whether the methodological requirements set out in the circular have been met (Bundesversicherungsamt, 2018). According to the Office's activity report for 2016, only eight out of 66 federally regulated health insurance funds had been unable to furnish evidence of the past cost-efficiency of their bonus programmes (Bundesversicherungsamt, 2016).

As with the motor insurers' telematics-based tariffs, it is probable that health-conscious individuals with low healthcare costs are most likely to participate in bonus programmes. This theory is confirmed, for example, by a study conducted by the Robert Koch Institute, which demonstrates that "A high level of health awareness and of health-conscious habits such as non-smoking and sporting activity are major influential factors in decisions to participate in a bonus programme. The findings indicate that the insured persons likeliest to take part in bonus programmes are generally those whose lifestyle is already very health-conscious and who would have participated in the preventive measures anyway; they are sucked into the programme, as it were" (Jordan, von der Lippe, Starker, Hoebel and Franke, 2015).

From a consumer's point of view it would be desirable if scientific evidence showed that the behavioural effect targeted by scoring schemes was actually being achieved. This applies both to the individual modes of behaviour or attributes that count towards the score and to the programme in its entirety. And a broader discussion is needed on the social desirability of bonus programmes that are predominantly embraced by people who are already healthy, even if these programmes turn out on closer inspection to be cost-effective, in other words if they make healthy people even healthier.

5. Discrimination

While the purpose of scoring is to distinguish between people and assess them individually, it should not result in anyone being unfairly disadvantaged or subjected to discrimination (see section B.II.2 above). We can start by asking which attributes are used for scoring. Some attributes, like ethnic origin and religion, are protected by the General Equal Treatment Act, and so from the outset their use is only permitted in certain circumstances. But there is also a need to discuss whether the inclusion of certain other attributes as scoring variables is liable to prove disadvantageous for particular consumers or groups of consumers to a socially unacceptable degree. The need for discussion seems greatest in the case of variables that would be difficult or impossible for a consumer to influence, which is why it seems unfair to base an assessment on these criteria (see also the findings of the consumer survey in Part D below).

A much-discussed aspect of credit scoring is geoscoreing (see also section B.III.3 above). Geoscoreing poses problems from a consumer perspective, because the scored person is deemed guilty by association for lax financial discipline of people in his neighbourhood, and the focus is no longer on individual payment practice. In this way it becomes impossible to influence one's own score. For this reason various privacy and consumer bodies have already branded this approach as discriminatory (e.g. Schaar, 2008; see also Britz, 2008).

Whereas one respondent credit reference agency ruled out geoscoreing altogether other than in exceptional cases, another stated that it took account, as a rule, of "payment experiences in the immediate residential environment".

Scoring on the basis of criteria such as age or sex may also be perceived as discriminatory because those attributes cannot be influenced. Two credit reference agencies addressed the variable of a person's sex when answering the question on potential discrimination problems. One agency excluded its use from the outset, while the other indicated that it only used that variable if it played a "razor-sharp" differentiating role in the model.

So two of the three credit reference agencies said they were prepared to distinguish by sex if the data indicated that it mattered. This practice may have discriminatory effects (see section E.III.5 below). It must be said, however that the third agency, which ruled out such distinction, ignored the possibility of indirect discrimination through the correlation of attributes such as sex or ethnic origin with other, permissible variables. Now it may be the case that empirical evidence reveals no actual influence of sex or ethnic origin on people's computed scores, but it is surprising that the problem of discrimination – be it direct or indirect – is categorically dismissed by some agencies. The response from one agency that mathematical statistical processes do not discriminate is an inadmissible oversimplification of the problem (see chapter B.II above and section E.III.5 below).



“Mathematical statistical processes do not discriminate.”

The discriminatory or prejudicial potential of health insurers' bonus programmes must also be considered in the context of the conditions in which the insurers operate. Statutory health insurance is based on the principle of solidarity. Briefly, this means that individuals' health risks are borne jointly by all insured persons. The assessment of contributions is based on the insured person's income and not on attributes such as age, sex or lifestyle choices. Entitlement to benefits is based on individual need, all insured persons having the same cover.

Participation in a bonus programme therefore has no influence on the assessment of an individual's contributions. Nevertheless, it may be argued that bonuses in the form of monetary rewards and benefits in kind lower the net cost of an individual's health insurance. The potential bonuses offered by some health insurance funds run to several hundred euros a year.⁷⁸ This raises the question whether bonus programmes are detrimen-

tal to particular groups of consumers because it is more difficult or impossible for them to share in such benefits and so lower their net insurance costs.

Bonus programmes are, in principle, open to all insured persons, as the survey respondents confirm. Yet not all measures seem to be equally accessible to all consumers. For example, sporting activities, participation in sports events or attendance at sports courses presuppose a certain degree of mobility and physical condition. Policyholders would be hampered or prevented from collecting bonus points in these ways if they were ill, physically impaired or elderly. Similar problems arise when bonus points are awarded for evidence of physical data and laboratory test results such as body-mass index or blood sugar level rather than for improvements in these indicators, for example. Structurally conditioned discriminatory effects are also conceivable, since a number of measures entail prior expenditure on the part of the insured person, as in the case of gym membership. Those who cannot afford such expenses may therefore be placed at a disadvantage.



“Behavioural tariffs may result in individual groups of insured persons exploiting them at the expense of people whose illnesses are not lifestyle-related. We therefore take a very critical view of these tariffs.”

Another subject for discussion is whether bonus programmes are used as a vehicle for indirect risk selectivity if those who are healthy anyway and those who are health-conscious are the main beneficiaries. The survey shows that some health insurance funds deliberately set out to appeal to health-conscious individuals. This may be interpreted as an attempt to recruit and retain the youngest and healthiest possible clientele.

⁷⁸ An overview of the maximum rewards in the bonus programmes of the various health insurance funds can be found at <https://www.gesetzlichekrankenkassen.de/bonuszahlung/bonuszahlung.html> (accessed on 1 October 2018).



“Through the potential for adverse selection, good health cover for persons who do not opt for behavioural tariffs may mean considerably increased premiums. The underlying reason is that good risks tend to choose the cheaper behavioural tariffs, which drives up contributions for people with non-behavioural policies because of the deterioration in the risk mix.”

This effect would surely be more marked if behavioural tariffs were available. Continuous linkage of individual contributions to health status and lifestyle would automatically be detrimental to consumers with health impairments.

The market survey also showed, however, that most of the statutory insurers attach great importance to the solidarity principle and want to protect it. Almost 50% of the respondents refer to jeopardisation of the solidarity principle or the danger of unfairness and discrimination as a disadvantage of behavioural tariffs. The solidarity principle was cited by 30% as the main argument against the introduction of a penalty system in which unhealthy lifestyles are penalised.

It is also advisable to examine telematics-based motor insurance tariffs too with a view to identifying any discriminatory effects arising from the use of particular variables. Behind the inclusion of scoring variables such as journey time, area of travel or road type lies the assumption that a journey made in certain basic conditions entails a higher accident risk. This, however, raises the question whether reliance on certain variables that a motorist can scarcely alter places certain groups of consumers at a disadvantage. A telematics user who lives in a city and therefore does a lot of urban driving will normally receive a worse score than a motorist whose circumstances cause him to do most of his driving across country. A night-shift worker whose driving mainly comprises night-time commuting will be assessed less favourably than someone who mainly drives

during the day. Such an approach seems to pose problems, because it cannot preclude discrimination against individual groups of consumers. Particular categories of consumers could be unfairly treated because of factors that they are scarcely able to influence. Moreover, as we have discussed above, it raises the question whether such comprehensive information is provided about the use of these specific variables that consumers are able to make decisions on the utility of a telematics-based option in their own particular circumstances.

6. Aggregation of data and inclusion of new consumer attributes

The companies taking part in the market survey were asked whether they could imagine using data from other areas of their customers' lives for scoring purposes. All of the respondents stated that they could not.

From a consumer perspective it would be critical, for example, if the scoring inputs of credit reference agencies were supplemented by variables from areas with no intuitive connection with creditworthiness, such as driving behaviour or health-consciousness. This would represent an aggregation of data from diverse areas of activity and start to resemble one of the super scores discussed in Part B above. In this respect it is interesting that all credit reference agencies in the survey stated that they were observing the trend which is already emerging abroad towards the inclusion of data from the Internet and social media in credit scoring (see chapter B.VIII above). One agency even spoke in this context of a considerable development potential. It is worth mentioning in this context that Schufa entered into a cooperative venture with the Hasso Plattner Institute back in 2012 to conduct basic research relating to the technical processing of public Web data. Under public pressure, however, the research institute subsequently terminated the cooperation agreement.⁷⁹

⁷⁹ See the press release from the Hasso Plattner Institute at <https://hpi.de/pressemitteilungen/2012/schufa-forschungsprojekt-gekuendigt.html>

Some health insurers showed signs of interest in extending their baseline data to include more health data. Some 37% of the respondents can imagine using data from the electronic patient file for bonus programmes or behavioural tariffs.

Only a very few of the respondent health insurers (five statutory funds) were in favour of the right to take account of their customers' unhealthy lifestyle choices as well as their healthy ones when devising bonus programmes or even when setting health insurance premiums. For a discussion of the issues surrounding bonus and penalty systems, see also Part D below.



“In the design of bonus programmes, it is entirely conceivable, in our view, that due consideration could be given to unhealthy lifestyle choices in the calculation of bonus entitlements. In a voluntary bonus programme, a better impact can surely be achieved with a bonus and penalty system than with a purely bonus-based system, since it would make the approach more holistic.”

Among the motor insurers the survey findings reveal no evident aspiration to extend the baseline data by adding new consumer attributes. In principle, the general view is that only variables relating to customers' driving behaviour are of interest for scoring purposes. Half of the providers estimate that public acceptance of a penalty system would be very low or non-existent. This assessment matches the findings of the consumer survey – see Part D below.

7. Supervision

Since the relevant companies can invoke trade secrecy in connection with their scoring methods and are therefore under no obligation to make these methods public (see section B.I.3 above), comprehensive supervisory practice is in the interests of consumers. Consumers with no insight into the details of scoring systems must be able to rely on the competent supervisory authorities being able to examine scoring algorithms critically and thoroughly and actually doing so (see chapter B.III above).

Credit reference agencies are regularly monitored by the supervisory authorities for data protection. Competence lies with the Data Protection Commissioner for the federal state in which the agency has its head office.

In the survey the three respondent agencies stated that they had disclosed their procedures to the competent *Land* data protection authorities. The responses raised the question of the actual nature of the supervision practised by these authorities, for whereas one agency stated that it had disclosed its scoring method to the Federal Data Protection Commissioner and the Data Protection Commissioners of all the *Länder*, another agency merely reported regular meetings with the competent Data Protection Commissioner. The Data Protection Commissioner for Hesse wrote in his activity report for 2014, “There is no area about which I receive more complaints than the activity of commercial credit reference agencies” (Der Hessische Datenschutzbeauftragte, 2014).

In his activity report for 2014/15, the Hamburg Commissioner for Data Protection and Freedom of Information wrote the following with regard to credit reference agencies: “For data protection authorities it is almost impossible to check for themselves whether the data that are used to calculate a probability value on the basis of a scientifically recognised mathematical statistical process are demonstrably significant as predictors of the probability of the relevant behaviour” (Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit, 2016). In May 2018, research carried out by reporters from the Bavarian public broadcaster Bayerischer Rundfunk revealed that checks on scoring systems by the data protection authorities were largely based on

specialists' reports that the credit reference agencies themselves had commissioned from universities and researchers (Kerler, Köppen, Schnuck and Zierer, 2018).

In the Bayerischer Rundfunk report, some leading voices from the field of data protection, such as Thilo Weichert, formerly Data Protection Commissioner for the federal state of Schleswig-Holstein, Peter Schaar, the former Federal Commissioner for Data Protection and the data protection authority of North Rhine-Westphalia, comment critically on this practice. They see conflicts of interest arising if the credit reference agencies pay for the specialists' reports, and they call for independent checks and provision of the requisite financial resources for that purpose. By contrast, the Office of the Hessian Data Protection Commissioner, which oversees several credit reference agencies, sees no problem in the approach, stating that the authorities could always commission additional specialists' reports of their own, although that had never yet been necessary.

As regards the supervision of telematics-based schemes offered by motor insurers, the survey paints a mixed picture. Five of the ten respondent insurers providing such schemes say that there is no state supervision at all of the use of telematics in car insurance. Four providers, on the other hand, refer to BaFin, the Federal Financial Supervisory Authority, as the competent supervisory body. The probable explanation for this divergence in the insurers' responses is that, while BaFin oversees insurance companies (for details, see section 4(1) of the Federal Financial Supervisory Authority Act (*Gesetz über die Bundesanstalt für Finanzdienstleistungsaufsicht*) and sections 320ff. of the Insurance Industry Supervision Act (*Versicherungsaufsichtsgesetz*)), it has not created a specific supervisory programme for telematics-based policies (see chapter E.IV below). In the market survey, none of the respondents referred in any detail to a review of scoring methods by a supervisory authority.

Under section 65a(3) of Book V of the German Social Code, health insurance funds must report to the competent supervisory authority at least once every three years confirming their compliance with the requirement to fund their bonus programmes from cost savings and efficiency gains. According to the Federal Insurance Office, "For the verification of cost-effectiveness, it is necessary to rely on the baseline data made available to the evaluator by the health insurance funds, since the Federal Insurance Office does not collect its own data." (Bundes-versicherungsamt, 2018). The Office, then, can only check whether the evaluation reports are complete and plausible.

The problem for consumers who have been promised health benefits is that it is not discernible from the evaluation reports addressed to the Federal Insurance Office whether participation in a bonus programme objectively benefits people's health. The Office can merely check whether bonus programmes meet the cost-effectiveness criterion but not whether the participants are behaving more health-consciously.

Another point that is open to consumer criticism is the divergence between the authorisation practices of the Federal Insurance Office, which is responsible for the federally regulated providers of statutory health insurance, and the regional supervisory authorities. This applies, for example, to the authorisation of actions for the common good as bonus-earning activities and to the use of fitness trackers. About 20% of the respondent insurers referred to this situation in their replies to the question on the possible need to adjust the legal and regulatory framework.



Public knowledge and acceptance of scoring



I. Preliminary study, 2017

Empirical information on public knowledge and acceptance of scoring and algorithmic decision-making is in short supply in Germany (see chapter B.VIII above). For this reason the SVRV conducted a representative population survey in cooperation with the infas Institute for Applied Social Sciences. This survey was preceded by a non-representative preliminary study. The findings of both studies are presented in this part of the report.

The preliminary study was conducted in Berlin in November 2017 to test parts of the questionnaire that had been developed for the representative survey. The study was conducted in a Berlin cinema auditorium with 91 participants (52% women), ranging in age from 17 to 75 (cf. Rebitschek, Gross, Brümmer, Gigerenzer and Wagner, 2018).

Paper questionnaires and pencils were distributed (Rebitschek, Gross, Brümmer, Gigerenzer and Wagner, 2018). The questions related, among other things, to knowledge of input variables used in established credit scoring systems, acceptance of scoring scenarios involving a combination of data relating to health and motor insurance, acceptance of direct reporting to the police of data from telematics equipment (penalty-notice scoring) and acceptance of comprehensive social scoring with data from various areas of people's lives. As an intervention portraying a fictitious social credit system, the *Nosedive* episode of the Netflix TV series *Black Mirror* was screened. In that film, every service people render and almost every social interaction is constantly being assessed; everyone has a personal score and can also find out anyone else's score in real time.

Acceptance of novel scoring systems combining variables from various scoring areas was confined to a minority before and after the film screening for all the scenarios in the questionnaire (see Annex V, table 1), social scoring being rejected by 92% of the participants before the film was shown and by 96% afterwards, while scoring on the basis of automatically generated penalty notices was approved by 31% before and 25% after the film. Almost 28% of the participants approved of scoring in at least one of the three scenarios. This percentage was even slightly higher before the film screening; this was to be expected, as most participants had presumably never given a great deal of thought to such scenarios before watching the film.

The intervention that gave the participants additional knowledge yielded the expected results, namely an increase in scepticism with growing awareness. The effects, admittedly, were limited, and so it may be assumed that the findings of a representative survey with no such intervention will carry greater weight. A detailed presentation of this study can be found in Annex V.

II. Representative survey, 2018

In order to obtain representative findings, the SVRV commissioned the infas Institute for Applied Social Sciences to conduct a computer-assisted telephone interview (CATI) survey in the period from the beginning of February to the end of April with an average duration of 22.5 minutes per interview. A total of 2,215 persons – 1,123 men and 1,092 women – took part; their ages ranged from 16 to 94. Since the degree of willingness to take part varied between sections of the population, these differences were smoothed out by weighting when the findings were analysed (infas, 2018). All of the percentage figures cited in the following paragraphs relate to the weighted results, but the absolute numbers of cases are not weighted. Because of rounding, the results in some cases do not always add up to exactly 100%.

Respondents were asked about their knowledge and acceptance of scoring in the three areas under examination in the present report, namely creditworthiness, third-party motor insurance and health. They were also asked about their acceptance of the correlation of scores from various areas. Here the respondents were to state whether they would sign up to a combination tariff in which they could cut their costs by making data about their health available to a motor insurer or vice versa. They were also asked about their acceptance of a social scoring system along the lines of the Chinese social credit system. To permit an empirical assessment of the significance of framing, the way in which, for example, a scoring-based behavioural insurance tariff is advertised, in the questions on motor and health insurance half of the respondents were presented with the consequences of scoring as a bonus, i.e. lower premiums for good scores, while the other half were presented with a penalty scenario, in which higher premiums were payable for poor scores.

The set of questions on motor insurance ended with the presentation of a fictitious scenario in which all cars had to be fitted with speed recorders, including compulsory direct transmission to the police of any infringement of the speed limit. The questions on health insurance ended with a fictitious scenario in which insured persons were required to record the number of paces they took each day and had to pay a share of their medical costs if their daily pace count was too low.

The structure and content of the questionnaire on the basis of which the telephone interviews were conducted is described in detail in Annex V. The questionnaire, including the precise wording of the questions asked in the telephone interviews, can be seen in infas (2018). The findings are also presented in descriptive detail in Annex V. The following paragraphs contain a more analytical presentation of the findings.

1. Analysis of the findings

1.1 Scoring: how much do people know?

Most of the established credit reference agencies include in their input variables past repayment behaviour, age and current loan agreement to assess a person's creditworthiness. Assets, activity in social networks, occupation and ethnicity, on the other hand, are not used for scoring purposes.

Table D.1 shows the respondents' replies to the question whether or not established credit reference agencies use the listed variables in the calculation of credit scores.

It emerges clearly that the level of knowledge about the use of attributes in credit scoring is moderate. Interestingly, the majority of respondents wrongly assumed that assets, occupation and ethnic origin are taken into account in credit scoring. In general terms it is observable that consumers overestimate the use of attributes, for the majority presumption is that six of the seven listed variables are used, whereas the real figure is only three.



Interim summary

The level of knowledge among the body of respondents about the use of attributes by established credit reference agencies to assess creditworthiness is moderate.

Table D.1

Knowledge about the use of creditworthiness attributes. Distribution of replies (in %) to the question whether the listed variable is used by credit reference agencies in the assessment of creditworthiness ('Yes') or not ('No'). Figures in green mean that the majority correctly identified the use or non-use of a variable; figures in red indicate that the majority were wrong about the use or non-use of the variable. The sample size was 2,215 (Source: infas, 2018).			
	YES	NO	NO REPLY / DON'T KNOW
Repayment behaviour	78	16	6
Age	71	23	7
Loan agreement	82	12	6
Assets	64	27	8
Social networks	30	57	13
Occupation	66	26	8
Ethnic origin	51	40	9

Table D.2

Frequency of communication of scores (Schufa: actual retrievals in %; demand for communication of scores in %)			
	YES	NO	NO REPLY / DON'T KNOW
Schufa score retrieved in the past 5 years	21	79	0
Acceptance of motor insurance scoring	44	55*	1
	daily / weekly / monthly only if major changes occur		
	27 17		
Acceptance of health scoring	46	54	0
	daily / weekly / monthly only if major changes occur		
	18		

* The response option here was *Gar nicht* (not at all) (Source: infas, 2018).

For a more detailed analysis we would have to know whether knowledge of the input variables that are used and not used by established credit reference agencies for credit scoring varies with particular factors, such as socio-economic status. It is conceivable, for instance, that respondents with an above-average formal education also have above-average knowledge about which consumer attributes are used in scoring. It is also possible that consumers in a particular age bracket know more about scoring because people in certain situations are more keenly aware of the implications of credit checks, for example if they are taking out a mortgage, which is normally done at an age between 30 and 50.⁸⁰ From among these and other potential influencing factors, such as monthly income, gender, whether one lives in a city or village, experience gained by looking up one's Schufa score in the past five years, we must identify those factors that correlate with a person's level of scoring-related knowledge.

Accordingly, regression analysis (see D.II.2 below) will be used to examine, for all the attributes listed in Table D.1, which factors determine the probability that respondents will correctly state whether an attribute is or is not actually used by credit reference agencies in the calculation of credit scores.

80 <https://www.sparkasse.de/themen/eigenheim-finanzieren/das-richtige-alter-beim-immobilienkauf.html>; accessed on 7 August 2018.

Prevalent practice for communicating a consumer's latest score ranges from not at all or only on request to the possibility of checking one's own score on an app at any time, which is available to motorists on telematics-based insurance tariffs. This wide dispersion of frequencies with which feedback is obtained through the communication of scores raises the question how often consumers normally retrieve their score (in the case of Schufa scores) and the frequency of feedback that consumers would welcome in the case of novel scoring models such as telematics-based driving scores or health scores.

It can be seen from Table D.2 that only about a fifth of the respondents say that they have found out their own Schufa score in the last five years by requesting a free copy of their personal credit record. Asked whether they were satisfied with the frequency with which telematics-based driving and health scores were communicated to them, about half of the respondents stated that they had no wish to learn their own score at all. Of the respondents who supported the principle of score notification, about a third expressed a wish to be informed only of major changes to their score; two thirds of those who advocated notification of scores wanted to be informed at least on a monthly basis, though some advocated weekly or even daily updates – regardless of whether or not their score had changed significantly.

i

Conclusion

A majority of the respondents are opposed to receiving automatic notification of their score. Almost half, however, would like to be informed of their score, although the desired frequency of communication varied. Current practice could conceivably be adapted by means of an opt-in arrangement so that automatic notification of a score would be generated whenever there was a major change in a consumer's score which would or might have implications for the consumer, for example if the change entailed an actual or potential drop into the next lower category.

1.2 Protection of personal attributes

As a rule, the calculation of scores involves recourse to consumers' personal data, such as repayment behaviour, vital parameters and exercise data. The respondents in the survey were asked about their acceptance of the collection and analysis of personal attributes from telematics devices

for the purposes of both motor and health insurance. The resulting acceptance and rejection rates are set out in Table D.3 below; the responses are listed in reverse order of acceptance rates, from majority rejection (less than 50% acceptance) to majority acceptance.

Table D.3

Acceptance or rejection of collection of attributes for a telematics-based motor or health insurance tariff (in %; listed from lowest to highest acceptance levels)*				
	VARIABLE	ACCEPTANCE (“Completely justified” or “Mostly justified”)	REJECTION (“Mostly unjustified” or “Not at all justified”)	NO REPLY / DON'T KNOW
Majority rejection (less than 50%)	Sleep ⁺	14	85	1
	Daytime/night-time driving [#]	17	79	3
	Walking ⁺	26	73	2
	Rural/urban driving [#]	30	69	1
	Weight ⁺	35	65	0
	Alcohol ⁺	43	56	1
	Acceleration [#]	45	53	3
Majority acceptance (more than 50%)	Smoking ⁺	58	41	0
	Speed [#]	62	37	1
	Cancer screening ⁺	63	37	0
	Using mobile phone to read or write messages [#]	66	33	1

*The table shows the survey findings for a bonus framing in which the fulfilment of requirements is rewarded

Input variables for a telematics-based motor insurance tariff.

+ Input variables for a telematics-based health insurance tariff (Sources: infas 2018 and our own calculations).

The collection and use of most of the potentially sensitive variables listed above for the calculation of scores was rejected by the majority of the respondents answering the bonus-framed questions relating to health and motor insurance. Only the variables smoking and cancer screening in the realm of health telematics the variables speed and mobile-phone use from the domain of vehicle telematics met with majority approval as input variables for the setting of insurance premiums.

INTERIM SUMMARY

On the whole, there is a low level of acceptance of the variables presented as input data for the driver and health scoring scenarios. This suggests that the majority of respondents are opposed to most of the listed behavioural variables being considered when insurance tariffs are set.

With the aid of regression analysis we shall examine which factors influence acceptance of the use of personal attributes for scoring purposes (see section D.II.2 below). The main hypothesis here is that the level of acceptance of the use of a particular variable for setting tariffs is partly dependent on the extent to which individuals are affected by that variable. If, for example, rural driving carries a bonus while urban driving lowers the motorist's score, because rural driving is associated with a lower accident risk than urban driving, it seems logical that acceptance of recourse to this variable will be lower among city-dwelling respondents. It also seems plausible that a person's state of health will be a factor in the acceptance of lifestyle-based health insurance premiums, as people in poor health might tend to reject such tariffs, while a person's normal mode of transport and place of residence might influence whether he or she accepts or rejects the use of particular variables in the calculation of motor insurance premiums.

1.3 Acceptance of new and potential developments in the scoring market

1.3.1 Would consumers use telematics-based driver and health scores?

About two thirds of the respondents reject novel and potential future scoring methods for motor and health insurance. Interestingly, this applies equally to both segments of the insurance market, and while the effects of bonus or penalty framing were evident, they were very small in numerical terms, especially with regard to health insurance (see Table D.4 below). Conversely, this means that about a third of the respondents can imagine the use of a behavioural tariff linked to scoring not only for motor insurance but also in the sensitive area of health. What makes the latter finding especially interesting is that the actual number of drivers whose motor insurance premiums are based on their driving

habits is well under a third of all policyholders (see Part C above on the market study), while the respondents are distinctly sceptical about the collection of potentially sensitive data (see subsection D.II.1.2 above), without which it would be altogether impossible to implement such a tariff system. This paradox is an indication that the intention to do something does not necessarily translate into action. On the other hand, the discrepancy between the high degree of scepticism shown by respondents towards the collection and use of data for a telematics-based tariff and the fact that about one third and more can imagine themselves using a telematics-based tariff – possibly for the potential savings – can also be explained by reference to the privacy paradox.

Table D.4

Acceptance of novel scores (in %). Respondents were asked to state whether or not they would use a telematics-based tariff. The answers relating to both motor and health insurance apply to the whole sample. and are divided into questions framed in terms of bonuses and those framed in terms of penalties.							
	YES			NO			NO REPLY / DON'T KNOW*
	Total	Bonus	Penalty	Total	Bonus	Penalty	
Motor insurance	36	43	30	63	57	69	1
Health insurance	34	36	32	65	64	67	1

Motor insurance: total sample size = 1,104, bonus scenario = 536, penalty scenario = 568; health insurance: total sample size = 1,111, bonus scenario = 570, penalty scenario 541 (Source: infas, 2018).

Examples of arguments against the use of a telematics-based motor insurance tariff



“Because a great deal of personal data would be collected that I do not want to disclose.”

(male respondent aged 40)

“I regard it as surveillance that interferes too much with my privacy.”

(female respondent aged 60)

“Because it takes snooping too far for my liking. [...] I expect solidarity from my car insurance. I would even pay more so as not to be snooped on.”

(male respondent aged 61.)

Examples of arguments for the use of a telematics-based motor insurance tariff



“It is about fairness at the end of the day. Everyone has the same opportunity for everything. Whoever keeps the rules should enjoy the benefits.”

(male respondent aged 33)

“Because I get advantages from it and contribute more to my safety and so protect others.”

(female respondent aged 27)

“Would only use it because I would benefit and am a careful driver.”

(female respondent aged 35)

Examples of arguments against the use of a telematics-based health insurance tariff



“Because I don’t want to let myself be monitored by the insurance fund. It is my privacy, and it is nobody’s business what I do and how. To my mind, it puts chronically sick or elderly people at an unfair disadvantage.”

(female respondent aged 57)

“I am all for the solidarity principle. Many people simply can’t do these things because of work obligations or illness. In those cases it would be pretty unfair.”

(male respondent aged 53)

“I would not like to record the readings or have them collected by someone. And everyone should be on a similar tariff, regardless of their personal constitution.”

(female respondent aged 63)

Examples of arguments for the use of a telematics-based health insurance tariff



“The solidarity principle of health insurance should be preserved for all insured persons; people who are ill through their own fault (as in the case of alcohol consumption) should be penalised.”

(male respondent aged 66.)

“Those who don’t do anything for their bodies should pay for it, and when people take everything into consideration, they should get a cheaper tariff (but always matched to their age).”

(female respondent aged 82)

“Because it is advantageous when you do something for your health, you get benefits from your health insurance too. It’s a win-win situation.”

(male respondent aged 29)

By means of a regression analysis, we shall identify the factors (see section D.II.2 below) on which acceptance of driving and health scores depends. In this case too, it seems logical to presume that acceptance of behavioural insurance tariffs will be highest among those who would benefit from them, e.g. particularly healthy or sporty individuals, who could expect to save money by taking out a lifestyle-linked policy.

1.3.2 What do people think of disclosing their individual scores?

Various motor insurance companies advertise the publication of their customers’ scores as a means of acknowledging exceptional driving performance. The positive effects of publishing the scores are supposed to be fully effective, if an intensified competition amongst the drivers fosters individual driving performance.

A disclosure of scores, however, is rejected by a clear majority, irrespective of whether the disclosure is done in a deliberate or compulsory fashion. As expected, the rejection rate is the highest in those areas that can be considered particularly sensitive, such as personal health. The framing of questions, that is whether the questions were asked in a bonus- or malus-framed fashion, does not affect rejection rates considerably. The respective results are presented in table D.5.



Conclusion

The disclosure of individual scores is rejected by a clear majority. This finding is applicable irrespective of whether the disclosure is done in a deliberate or compulsory fashion.

Table D.5

Consumers’ willingness for their own scores to be published on a voluntary or compulsory basis (in %). The figures in brackets relate to compulsory publication.							
	YES			NO			NO REPLY / DON’T KNOW*
	Total	Bonus	Penalty	Total	Bonus	Penalty	
Credit score	12 (7)	--	--	87 (92)	--	--	1 (1)
Driving score	12 (10)	13 (13)	11 (8)	87 (90)	86 (87)	88 (92)	1 (0)
Health score	5 (5)	4 (6)	7 (5)	95 (94)	96 (94)	93 (95)	0 (0)

Credit score: total sample size = 2,215; driving score: total sample size = 1,104, bonus scenario = 536, penalty scenario = 568; health score: total sample size = 1,111, bonus scenario = 570, penalty scenario 541 (Source: infas, 2018).

1.3.3 What do people think of super scores and score-based penalisation?

The participants in the survey were also asked whether they approved of super scores, i.e. the correlation of scores from various areas, and of score-based sanctions in hypothetical scenarios. The scenarios covered: (1) the correlation of two scores (driving and health), (2) a system of automatic penalisation in the event of errant behaviour through a communication channel with 'higher authorities', i.e. the police or a doctor, and (3) a stylised version of a social credit system, based on a composite score drawn from more areas of a person's life than those used in scenario (1).

Correlation of data held by motor and health insurers met with the approval of almost a quarter to almost a third of the respondents. The automatic transfer to the police of logged infringements of speed limits was also acceptable to more than a quarter of the respondents.

This acceptance rate fell sharply in the hypothetical case of a fitness tracking device registering a failure to take the required daily number of paces, thereby making its owner liable for a share of his medical costs. A social credit system met with similarly low acceptance.

The rejection of penalties for lack of exercise could be due to the fact that the pace count touches on an area which is part of the sensitive domain of health in the widest sense and that people respond considerably more antipathetically to scoring if they are made aware of the potentially adverse consequences of scoring – in this case monetary consequences, namely having to meet a share of the medical costs. In the case of an all-embracing collection and aggregation of data along the lines of the social credit system, this appears to be very much unwanted. A summary of the findings is set out in Table D.6 below.

Table D.6

Acceptance of super scores and score-based sanctions in the whole sample of 2,215 respondents (in %)				
		YES	NO	NO REPLY / DON'T KNOW
Score aggregation	Combined tariff: motor and health insurance*	24	76	0
	Combined tariff: health and motor insurance**	31	68	1
Behaviour penalisation	Speeding: report transmitted to the police	28	72	0
	Lack of exercise: contribution to medical costs	8	91	0
Social credit system	Driving, health, invoice settlement, respectful conduct	9	91	0

* The respondents were divided into two groups. This group was initially presented with a bonus system forming part of a motor insurance policy. The respondents were then asked whether they could imagine having this score linked with a health score in order to obtain a reduction in their premiums.).

** Vice versa.

The 'Yes' figure includes the responses 'Yes, certainly' and 'Yes, possibly'; the 'No' figure encompasses those who responded 'No, probably not' or 'No, certainly not' (Source: infas, 2018).

Table D.7

Acceptance of super scores and score-based sanctions, broken down into questions framed in terms of bonuses and those framed in terms of penalties (in %).				
		YES	NO	NO REPLY / DON'T KNOW
Score aggregation	Combined tariff: motor and health insurance*	M: 24 (24)	M: 76 (76)	M: 0 (0)
	Combined tariff: health and motor insurance**	H: 32 (30)	H: 68 (69)	H: 0 (1)
Behaviour penalisation	Speeding: report transmitted to the police	M: 31 (24)	M: 69 (76)	M: 0 (0)
	Lack of exercise: contribution to medical costs	H: 6 (10)	H: 93 (89)	H: 0 (0)
Social credit system	Driving, health, invoice settlement, respectful conduct	M: 10 (10)	M: 90 (90)	M: 0 (0)
		H: 7 (7)	H: 92 (92)	H: 0 (0)

* / ** See Table D.6. The figures without brackets show the percentages of respondents to the questions framed in terms of bonuses, the figures in brackets to the percentages of respondents to the questions framed in terms of penalties. The 'Yes' figure includes the responses 'Yes, certainly' and 'Yes, possibly'; the 'No' figure encompasses those who responded 'No, probably not' or 'No, certainly not' (Source: infas, 2018).

Examples of arguments against the use of a combined tariff (motor and health insurance)



“The one thing must not have anything to do with the other; consider risk factors separately for each insurance type.”

(male respondent aged 56)

“There is something a bit scary about an insurance firm getting all my health data. You basically don't know where it will actually end up.”

(female respondent aged 36)

“I utterly oppose all this recording of data and details of our personal lives.”

(male respondent aged 63)

Examples of arguments for the use of a combined tariff (motor and health insurance)



“I'm always one for innovative products and services if they offer a certain added value. This kind of thing would harness synergetic effects that could possibly have great potential. Yet I am not entirely sure whether it maybe goes too far in this case.”

(male respondent aged 50)

“Because I do think it's a good thing if we all show greater awareness on the road and in the way we live.”

(female respondent aged 52)

“It gives you benefits like possibly saving you money. For me personally it would be good, because I meet the criteria.”

(female respondent aged 56)

The differences in the acceptance rates for super scores and score-based sanctions between the subsample whose questions focused on bonuses and the one whose questions were framed in terms of penalties can at best be considered marginal, both in relation to the whole group and between the two subsamples (see Table D.7 above).

In a more in-depth examination, we shall use regression analysis to identify the factors associated with acceptance or rejection of super scores and score-based sanctions (see section D.II.2 below).

i

CONCLUSION

A large majority, in some cases the vast majority, of respondents reject the correlation of scores. In particular, the idea of monetary sanctions for unhealthy lifestyles and the aggregation of scores from all areas of people's lives are overwhelmingly rejected.

1.4 The significance of framing in terms of bonuses or penalties

Telematics-based motor insurance tariffs are generally offered nowadays in the form of a bonus system, that is to say as a scheme in which consumers opting for the tariff obtain discounts on the basic premium if they drive well. The gains to be made depend on the amount of the basic premium – the higher it is in relation to the non-telematics-based tariff, the more difficult it will be to make net savings by taking up the telematics option. It is, moreover, conceivable that insurers will also incorporate penalty elements into their future tariff systems to create even stronger motivation for motorists to drive safely. Framing, in other words the question whether scoring is presented to consumers in the context of insurance tariffs as a potential source of gains or losses, may therefore be significant and may considerably influence consumer acceptance, not only of data collection as such but also of scoring in general.

The survey findings show that the way in which questions are framed – whether they focus on bonuses or penalties – sometimes makes a significant difference. It emerges clearly that framing alone may cause approval rates to rise or fall. The difference between approval ratings resulting from bonus-framed and penalty-framed questions concerning the actual or potential use of variables as inputs for the calculation of novel telematics-based tariffs is shown in Table D.8 below. The difference between bonus-based framing and penalty-based framing is particularly high for the use of speeding as a variable, whereas the approval rate for the inclusion of smoking is not influenced by the framing of the question.

It may be assumed, for instance, that insurers offering telematics-based tariffs will tend to advertise such tariffs in a favourable light, in other words along the lines of the bonus-based question framing employed in the survey. This may mean that consumers will be less aware of the potential for losses inherent in the bonus system or of the danger of abuses in the collection of their personal data. It is a fact that behaviour which does not conform to the applicable requirements can lead to the deduction of points at any time, which reduces the bonus. It is also unclear at the present time how unscored basic tariffs will develop in the long term and whether penalty elements will find their way into telematics-based schemes in future.

i

Conclusion

Acceptance rates for the inclusion of a variable as a pricing factor in a motor or health insurance policy vary depending on whether the question is framed with a focus on bonuses or on penalties.

Table D.8

Differences (in percentage points) in the acceptance of the collection of data as input variables and of novel scores depending on whether the question is framed in terms of bonuses or in terms of penalties			
		ÄNDERUNG*	
Collection of input variables*	Speeding	↓	-31
	Walking	↓	-18
	Cancer screening	↓	-17
	Rural/urban driving	↓	-15
	Daytime/night-time driving	↓	-14
	Sleep	↓	-9
	Alcohol	↓	-8
	Weight	↓	-6
	Smoking	↔	0
	Acceleration	↑	+4
	Mobile-phone use	↑	+12
Novel and potential future scores**	Driving scores	↓	-6
	Health scores	↓	-2

* The figures shown here represent the variation in approval rates, i. e. the percentages of respondents choosing the reply option "Entirely justified" or "Mostly justified".

2. Multivariate regression analyses: presentation and discussion of findings

The presentation of findings in this section is confined to a simplified outline of the main results of the multivariate regression analyses. The analyses are intended to

identify the factors that are associated with knowledge of the consumer variables used by credit reference agencies or with acceptance of novel telematics-based tariffs in motor and health insurance. A detailed presentation of the regression analyses, including the statistical parameters, can be found in infas (2018). The findings are adjusted by means of the Bonferroni-Holm sequentially rejective multiple test procedure (Holm, 1979).

i

Factors influencing knowledge of the variables used in credit scoring (Table D.1 above and Table 2 in infas, 2018)

Table 2 in infas (2018) shows the number of correctly identified credit-scoring inputs. With the aid of linear regression, infas examined the influence of certain socio-demographic attributes and attitudes of respondents on the number of correct answers they gave regarding the collection of consumer data by credit reference agencies. The dependent variable is modelled as a continuous variable with the number of correctly identified consumer attributes ranging from 0 (no attribute correctly identified) to 7 (all attributes correctly identified). From the regression coefficient – provided that it is significant – it is possible to elicit how many more attributes a given group of persons identify correctly on average compared with a reference group.

Example: The regression coefficient of -0.15 for the variable ‘equivalised household income: below average’ tells us that respondents with a below-average equivalised household income provide an average of 0.15 fewer correct answers than respondents whose equivalised household income is average.

For all subsequent analyses (logistic regressions) from Tables 3 to 8 in infas (2018), it may be said that the differential between the regression coefficient and 1 expresses the strength of the effect. Expressed in percentage terms, this means, for example, that a coefficient of 1.78 indicates that the odds on the occurrence of an event are 78% higher than in the case of the reference group. A regression coefficient of less than 1 means that the probability of a correct answer in this group is lower than in the reference group.

Example: The regression coefficient of 1.8 for the variable ‘age: 16–34 years’ (Table 1 in infas, 2018) means in percentage terms, if the coefficient is significant, that people aged 16 to 34 are 80% more likely than the reference group (aged 65 and over) to answer correctly the question whether final non-payment of invoices is used as an input variable by many credit reference agencies.

2.1 Knowledge relating to the use of variables in credit scoring

People's level of knowledge about credit scoring is associated primarily with their age and whether they have obtained their Schufa credit score in the last five years. The total number of correct answers given by respondents who have requested their Schufa score in the last five years is slightly higher on average than the number of correct answers from those who have not asked for their Schufa score during that time. Respondents in the age brackets from 16 to 64 also average half a correct answer more than their counterparts aged 65 and over. As regards the individual input variables that are used for credit scoring, it emerges that the probability of a correct answer regarding the variables 'Current loan agreement', 'Data from social networks' and 'Ethnic origin' is higher among younger respondents, i.e. those in the 16–34 and 35–49 age brackets, than among the respondents aged 65 and over. The probability of a correct answer to the question whether data from social networks are used in assessing creditworthiness is also higher among respondents who have obtained their Schufa score in the past five years than among those who have not. Among respondents in the 16–34 age bracket the probability of a correct answer to the question whether ethnic origin features as an input variable in credit scoring is higher than among their older counterparts in the 65+ age group.

Overall, the findings indicate that respondents are most likely to know which variables are and are not used as inputs in creditworthiness assessment if they themselves have made active efforts to obtain their personal credit record or if they belong to the age group that is most likely, for example, to go through the process of applying for a mortgage and have to concern themselves with their credit score in that context. Since there was no observable systematic correlation between general formal education and correct identification of the variables used by credit reference agencies, the relevant knowledge is evidently not sufficiently addressed through formal education. It therefore appears that it would be useful to impart specific knowledge of scoring – what we might call promoting scoring literacy – both in the formal education framework and by other means.

2.2 Acceptance of data collection by in-vehicle telematics (Tables D.3 and D.4 above; Tables 3 and 4 in infas, 2018)

The age of respondents as well as the size of the settlement where they live in terms of population play a part in acceptance of rural driving as a factor that counts towards an advertised bonus in the form of a reduction in the cost of insurance premiums. Among older people, i.e. those aged 65 and over, the probability of approval of this input variable is higher than among people in the middle age bracket of 35 to 39 years. The probability of acceptance of rural driving as a positive input variable for policy pricing is also higher on average for people who live in a town with a population of 20,000 to 100,000 than for people who live in a city with a population in excess of 100,000. Here there are signs of a knock-on effect within the group of people who live in towns, i.e. urban settlements with fewer than 100,000 inhabitants: the smaller the town, the higher is the level of acceptance of schemes promising bonus points for rural driving.

In addition, the findings show that the probability of acceptance of speed as an input variable is higher among younger respondents (16–34 years of age) compared with those aged 65 and over when the question is framed with the focus on penalties.

2.3 Potential uptake of telematics-based tariffs and combined telematics-based tariffs (Tables D.6 and D.7 above; Table 7 in infas, 2018)

With regard to in-vehicle telematics, it is apparent that the potential uptake of such a tariff system is highly dependent on the context and the prospective consequences. Those to whom the telematics-based tariff has been presented as a bonus scheme can more readily imagine signing up for such a tariff than those to whom the potential penalties have been emphasised. This may mean that motor insurance tariffs based on telematics do not appeal to some groups of people unless the possibility of a price saving exists. The framing of the telematics option thus appears to play a considerable part in terms of the potential public uptake of offers based on scoring.

What is being seen in the realm of health telematics is that people with an above-average internal locus of control (“My future is in my own hands”; “If I try hard, I will succeed”) find it easier to imagine being on a telematics-based health insurance tariff than those whose internal locus of control is below average. This means that those who believe they can control their own lives and that their own actions will lead to success can more easily imagine themselves with a telematics-based health insurance policy.

In addition, the most price-conscious consumers can most readily imagine themselves on a combined tariff, in other words a tariff in which they would pay lower motor insurance premiums in return for making health-related data available to the insurer.

2.4 Acceptance of telematics with penalisation (Table 8 in infas, 2018)

The recording of pace counts with automatic penalisation in the form of billing for a share of medical costs in the event of insufficient exercise is more frequently approved by people who take part in sporting activity more than once a week than by those who are not so active and engage in sporting activity once a week or less. This suggests that people who have no reason to expect personal penalisation on any particular grounds, such as lack of exercise, are more inclined to accept a system that entails scoring-based penalisation on those grounds.

Speed recording with automatic penalisation, namely notification of speeding offences to the police, tends to be more acceptable to those who have no penalty points on their licence in the national register at Flensburg than to those who do. To put it in more general terms, acceptance of scoring seems to be higher among people who have not come to the attention of the authorities for infringements in the area in which they are being or will potentially be scored. We could therefore speculate that people who already tend to drive carefully, for instance by keeping to the speed limit and accelerating gently, and who are thus more likely to obtain bonuses, will more readily accept a telematics-based motor insurance tariff than those who do not match this description. This type of scoring also meets with greater approval from respondents who mainly use public transport than those whose main form of transport is a private car. These findings show that this form of scoring is more widely approved by people who will presumably be less affected by it and by the associated sanctions.

Generally speaking, we can therefore say that the hypothetical scores described here tend to be accepted as long as the respondents themselves have little or no reason to expect adverse consequences from such a scoring system.

3. Population survey findings: general summary and conclusions

3.1 On the transparency and comprehensibility of scoring

A majority of the respondents are against receiving notifications of their score, while the others (almost half of all respondents) would like to be informed in principle, albeit with variations in their preferred notification frequency. Current practice could conceivably be adapted by means of an opt-in arrangement so that automatic notification of a score would be generated whenever there was a major change in a consumer's score which would or might have implications for the consumer, for example if the change entailed an actual or potential drop into the next lower category.

Acceptance rates for input variables, in other words whether a the use of particular variable in the calculation of motor or health insurance premiums is deemed to be justified, also differ depending on whether the variable is presented in connection with bonus or a penalty. To enable consumers to make informed decisions on their chances of gains and losses, it would be desirable for them to receive information from their insurers on average bonus amounts and, where relevant, on possible losses.

3.2 On knowledge of scoring and scoring literacy

Across the whole spectrum of respondents, knowledge about the use of attributes in credit scoring is moderate.

The findings show in detail that the level of knowledge about the attributes used in credit scoring depends both on a person's age and on whether he or she has obtained a personal credit record in the past five years. Formal education alone plainly does not seem to have any significant influence in this respect. Besides fostering scoring-related knowledge, it would appear to be beneficial to impart skills that are specific to scoring. The use of personal credit records that are obtainable on request, for example, seems to be a useful and comparatively inexpensive means to this end. This resource should be used expressly to give consumers important information in easily comprehensible terms about input variables and, where appropriate, their relative weighting so that they can engage in scoring processes in an informed manner.

3.3 On non-telematic options

On the whole, acceptance of the attributes proposed as input variables for the driving and health scoring systems we have presented is low. This suggests that the majority of respondents are opposed to the inclusion of most of the behavioural and situational variables presented above in new insurance tariffs, maybe not least because some of them relate to sensitive areas such as personal health.

The detailed findings show that, for some groups of people, acceptance of the collection of data on personal attributes and acceptance of telematics-based schemes in general are dependent on the extent to which they affect these people and on their personal circumstances:

- Public transport users support automatic reporting to the police of motorists' speed-limit infringements.
- Those who engage in sporting activity anyway tend to support telematics-based health insurance.
- Motorists who have penalty points on their licence tend to be opposed to in-vehicle telematics.

Although higher acceptance rates were recorded for specific groups of people, acceptance of the collection of personal data as input variables for telematics-based insurance tariffs is, as a general rule, low. What is more, partly because the factors that could probably lead to higher acceptance cannot easily be influenced by everyone, such as whether people use public transport or enjoy unhindered physical mobility, the preservation of non-telematic options in the range of insurance policies would be welcome.

In terms of fairness it may also be regarded as problematic that the range of scoring-based products involving telematics meets with higher acceptance among people who are highly price-conscious. Although it is essentially gratifying when scoring enables consumers to obtain cheaper insurance premiums, an implicit compulsion to use telematics and hence to disclose personal data simply to avoid financial predicaments would be very undesirable from a consumer perspective. For these reasons too, consideration should be given to offering consumers a permanently available non-telematic (low-disclosure) option in order to guarantee genuine freedom of choice.

3.4 On super scores

A large majority of the respondents reject any publication of scores that characterise them. This applies both to publication on a voluntary basis and to a general publication requirement. The respondents also reject aggregation of scores from various areas of people's lives, in some cases by a very large majority.

The greatest opposition is encountered by the idea of monetary penalties for unhealthy lifestyles in the realm of health insurance and by the mooted creation of composite scores covering every area of people's lives.



The legal framework for scoring



The SVRV defines scoring as the assignment of a numerical value to a person for the purpose of predicting or guiding that person's behaviour. That numerical value is normally determined by applying an algorithmic procedure to a broad set of baseline data (see chapter A.I above). Although there are certainly legal provisions governing scoring defined in this way, these provisions are scattered among a wide variety of legal instruments. There is no codified regulation of scoring, let alone a 'Scoring Act'.

Thematically limited legal requirements for scores are derived from various sets of provisions. Depending on who undertakes scoring in the above sense and on the purpose for which it is done, who are being scored, on what aspect of their lives they are being scored and what legal or practical implications the computed score will have, various areas of the law and legal provisions are applicable in determining whether particular scoring operations are lawful. At this point it should be stated for the avoidance of doubt that, while scoring is explicitly regulated in section 31 of the Federal Data Protection Act, which has influenced the understanding of the term in the public debate, scoring within the meaning of that provision covers only some aspects of the phenomenon under examination in this report (for details, see section E.I.3 below). Scoring cuts across established legal fields and has not yet been the subject of legislative action as a specific phenomenon in need of regulation.

In the following sections we shall examine key interfaces between scoring and the legal system. The initial focus will be on scoring as a data-processing operation, and then requirements for scoring in specific sectors will be described. This will necessarily be done by means of examples, given the diversity of guises in which the concept of scoring appears. Following on from this description, we shall discuss the potential of current law for resolving general scoring-related problems concerning mathematical and statistical quality, transparency and non-discrimination. We shall conclude with a brief look at supervisory structures that could be harnessed for the enforcement of more stringent requirements for scoring (see the recommendations for action set out in Part F below).

I. The basis in data privacy law

Article 22 of the General Data Protection Regulation (GDPR) and section 31 of the Federal Data Protection Act come closest to performing the function of a set of rules for the regulation of scoring. Neither of these instruments covers scoring as defined in this report, at least not in its entirety. The definition of profiling in the General Data Protection Regulation does not contribute to a targeted regulation of scoring, since it is not accompanied by a clear definition of legal consequences (see section 1 below).

As a data-processing operation, scoring must, of course, satisfy the general requirements of data privacy legislation. On the challenges that scoring poses to existing conventions, namely the declaration of consent to the use of personal data (see section B.VIII.2 above), the principle that personal data may be used only for the purpose for which they were collected and the principle of data minimisation, see the SVRV working paper *Verbraucher-Scoring aus Sicht des Datenschutzrechts* (Domurath and Neubeck, 2018), which supplements the present report.

It is not yet possible to estimate what impact the principles for the processing of personal data laid down in Article 5(1) GDPR will have on scoring. Be that as it may, the potential of these principles to set standards (Frenzel, 2018, on Article 5 GDPR, points 55–56) cannot be dismissed as minimal from the outset, a fact highlighted, for example, by studies on the principle of fair data processing within the meaning of Article 5(1)(a) GDPR (Maxwell, 2015; Hacker, 2017).⁸¹

1. Profiling (Article 4(4) GDPR)

Article 4 of the GDPR defines numerous basic concepts in data privacy law in the form of a catalogue. Article 4(4) GDPR contains a definition of profiling.

i

Article 4 GDPR Definitions.

For the purposes of this Regulation:

(...)

(4) ‘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;”

1.1 Profiling as an activity with no legal consequences

Scoring as defined in this report fits comfortably under the heading of profiling (Domurath and Neubeck, 2018; Schild, 2018, on Article 4 GDPR, point 64; Martini, 2018, on Article 22 GDPR, point 7). The only snag is that the definition of profiling in the General Data Protection Regulation remains inconsequential. The activity of profiling does not have any particular legal consequences (Veil, 2008, on Article 4(4) GDPR, point 1, and on Article 22 GDPR, point 4).

⁸¹ In the widely discussed *Google Spain* judgment delivered by the Court of Justice of the European Union on 13 May 2014 – Case No C-131/12 [EU:C:2014:317] – the Court deduced from Article 6(1)(c) to (e) of the Data Protection Directive (Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data – the forerunner instrument to the General Data Protection Regulation) – that it constitutes inadmissible processing if an internet search engine displays certain results in the context of a name search and so establishes a ‘right to be forgotten’ (Schantz, 2018, on Article 5 GDPR, point 2).

Although the term ‘profiling’ recurs in several provisions of the General Data Protection Regulation,⁸² in each of those provisions the term could be omitted without altering the regulatory scope of the provision (Veil, 2018, on Article 4(4) GDPR, points 3–4, and on Article 22 GDPR, point 53).⁸³

In addition, several recitals of the General Data Protection Regulation refer to profiling.⁸⁴ Recitals are not “second-class law” but an integral part of the relevant legislative act of the Union, a part for which the primary law of the EU provides (see the second paragraph of Article 296 TFEU). Such recitals, however, do not establish any rights or obligations, so it would be misleading to refer to this reference to profiling as a provision.

The reason for this legal position is that it proved impossible, in the legislative process for the General Data Protection Regulation, to reach agreement on the legal consequences of profiling. At issue was not only what the consequences of profiling activity should be (Veil, 2018, of Article 4(4) GDPR, points 9ff; cf. also WP 29, 2013), but also the preceding question as to which characteristics and circumstances should be emphasised in order to give the profiling phenomenon some contours in the first place, to which various answers were put forward in the course of negotiations on the General Data Protection Regulation (Veil, 2018, on Article 4(4) GDPR, points 5–6). It may be that the legislators’ decision not to attach any legal consequences to profiling was ultimately the key to the formal consensus reflected in Article 4(4) GDPR (SVRV, 2016).

1.2 Profiling as a weighting criterion

“The fact that profiling is nevertheless defined” – i. e. in spite of its lack of legal consequences – in the General Data Protection Regulation “has a purely political significance and is intended to signal that the lawmakers at least recognised the challenges associated with various forms of profiling” (Veil, 2018, Article 4(4) GDPR, point 1). On the basis of this “statement function” of a definition of profiling, it also seems fair to assume that the presence of data processing in the form of profiling will have the effect of placing a weight on the scale wherever a balance has to be struck between the interests of data processors and data subjects. In many places, the General Data Protection Regulation regulates the admissibility of data-processing operations and the conditions for the exercise of data subjects’ rights by means of general clauses (Buchner, 2017; on the reasons for this regulatory model in data privacy law, see Petersen, 2000, and Hoffmann-Riem, 1998).

General clauses are particularly amenable to the incorporation of numerous evaluation criteria that must be determined situationally and weighted. These clauses require the reconciliation of conflicting interests and objectives of the common good but do not prescribe the outcome of that reconciliation. General clauses therefore guide the application of the law to a comparatively low degree. This applies to both the direction of data controllers (Article 4(7) GDPR), who seek to shape their data-processing operations and structures in compliance with the law, and the programming of oversight activity on the part of supervisory authorities and courts, which ultimately have to decide on the legality of such data-processing arrangements.

If a data-processing operation is definable as profiling within the meaning of Article 4(7) GDPR, it is legitimate to lend considerable weight to the data subject’s need for protection in a legally prescribed balancing of interests. This takes account of the need to protect the data subject, which recital 71 of the General Data Protection

⁸² Siehe Art. 13 Abs. 2 Buchst. f; Art. 14 Abs. 2 Buchst. g; Art. 15 Abs. 1 Buchst. h; Art. 21 Abs. 1 Satz 1 Halbsatz 2, Abs. 2 Halbsatz 2; Art. 35 Abs. 3 Buchst. a; Art. 47 Abs. 2 Buchst. e; Art. 70 Abs. 1 Buchst. f DSGVO.

⁸³ Soweit ersichtlich einzige Ausnahme: Der Europäische Datenschutzausschuss (Art. 68 ff. DSGVO) hat unter anderem die Aufgabe, Leitlinien usw. zur näheren Bestimmung der Kriterien und Bedingungen für die auf Profiling beruhenden Entscheidungen gemäß Art. 22 Abs. 2 DSGVO bereitzustellen (Art. 70 Satz 2 Buchstabe f DSGVO); nicht jede der von Art. 22 Abs. 2 DSGVO erfassten automatisierten Einzelfallentscheidungen beruht aber auf Profiling-Vorgängen. Ein Grund für die nach dem Textbefund der Vorschrift gegebene Einschränkung des Aufgabenbereichs des Ausschusses (die man nicht als Redaktionsversehen abtun kann, s. EG 72 Satz 2 DSGVO) ist nicht erkennbar.

⁸⁴ Siehe vor allem EG 71 DSGVO, weiter EG 60 Satz 3, EG 63 Satz 3, EG 70 Satz 1, EG 72 Satz 1 und 2, EG 91 Satz 2 DSGVO.

Regulation explicitly recognises, in the application of the law. This may be illustrated by Article 6(1)(f) GDPR, which states that processing is lawful if it “is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data”.

In the balancing of interests that is required by that provision, considerable weight must be attached to the data subject’s interest in not having his or her data processed for profiling purposes (Buchner, 2018, on Article 4(4) GDPR, point 8).

1.3 Potential for the regulation of scoring by Article 4(4) GDPR

The conclusion to be drawn is that the definition of profiling in Article 4(4) GDPR does not amount to a regulation of that phenomenon but only lays emphasis on the social significance of that data-processing operation. A materially appropriate regulation of scoring cannot be built on that foundation. Although it is entirely possible for the Court of Justice of the European Union and national courts to set additional generalised standards supplementing the current legal rules, that is not a particularly realistic scenario, given the relatively scant intervention of the judiciary so far to flesh out the provisions of data privacy law.

2. Automated individual decision-making (Article 22 GDPR)

i

Article 22 GDPR

Automated individual decision-making, including profiling

- (1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
- (2) Paragraph 1 shall not apply if the decision:
 - a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
 - b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests; or
 - c) is based on the data subject’s explicit consent.
- (3) In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
- (4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests are in place.

2.1 The structure and legal consequences of the provision

Article 22 GDPR lays down criteria for the legality of automated decisions in individual cases. The provision displays an impressive legislative complexity: it identifies a basic form of decision (paragraph 1), exceptions to that form (paragraph 2), then an exception to those exceptions (first clause of paragraph 4) and lastly an exception to the latter (second clause of paragraph 4).

Paragraph 1 of the provision lays down the principle that data subjects have a right not to be subject to a decision if it produces legal effects concerning them or similarly affects them to a significant degree. Paragraph 2 provides for three exceptions to this rule, permitting automated individual decisions if their purpose is the performance of a contract, if they are based on a law or if the data subject has given his or her consent. In section 37 of the Federal Data Protection Act, the German legislature created such a legal basis for further restriction of the data subject's right. Paragraph 4 of Article 22 GDPR restricts the trio of exceptions in certain cases in which the automated decision is based on special categories of personal data, i. e. those referred to in Article 9(1) GDPR. If an automated individual decision is admissible under this set of rules, Article 22 GDPR makes an additional stipulation that suitable measures must be laid down to safeguard the data subject's rights and freedoms and legitimate interests (paragraphs 2(b), 3 and 4).

The GDPR establishes numerous consequences that apply in the event of an automated individual decision within the meaning of Article 22 GDPR (these are summarised in Veil, 2018, on Article 22 GDPR, points 16ff.). Among these consequences are specific information obligations prior to the processing of data (Article 13(2)(f) GDPR where data have been collected from the data subject and Article 14(2)(g) where data have been collected from other persons) and information rights after data processing (Article 15(1)(h) GDPR); for more details see section E.III.1 below. To these may be added "system- and process-related obligations" (Dreyer and Schulz, 2018, pp. 32ff.), which include in particular the obligation to carry out a data protection impact assessment (Article 35 GDPR), to enact binding corporate rules (Article 47(2)(e)) and to designate a data protection officer (Article 37 GDPR, taken in conjunction with the second sentence of section 38(1) of the Federal Data Protection Act and Article 35 GDPR).

2.2 The 'decision' as a key criterion in Article 22 GDPR

Article 22 GDPR cannot fulfil its potential for the regulation of scoring unless scoring meets the condition of being "a decision based solely on automatic processing" which "produces legal effects" concerning the data subject or "similarly significantly affects him or her". In the following paragraphs we shall use the term 'the decision' within the meaning of Article 22 GDPR to refer to this phenomenon.

In which cases a 'decision' within the meaning of Article 22 GDPR is encountered is far from self-evident (*Gesellschaft für Informatik*, 2018). This uncertainty is well illustrated by the example of scoring itself. The assignment of number x to person P may be described as a decision by analogy with that of the judges at a gymnastics competition or with marks awarded for schoolwork. If every assignment of this type, however, were regarded as a decision within the meaning of Article 22 GDPR, the applicability of that provision would have to be virtually ubiquitous. The character of the decision would forfeit any distinctiveness and could no longer perform the function of narrowing down the definition of an unlawful situation.

By using the definitional element of a 'decision', Article 22 GDPR distinguishes between interactions in which one person takes a decision about or for another on the one hand and structurally parallel interactions between a machine and a person on the other hand; only in the latter case does the General Data Protection Regulation establish the special regime described above with Article 22 at its heart. The legislators, however, lacked a clear concept of what the problematic element of these machine decisions they were regulating was actually meant to be (Dammann, 2016; Veil, 2018, on Article 22 GDPR, point 3), the element that was supposed to justify bringing them under legislative control. Against this backdrop, the characterisation of the provision as the "expression of a vague general disquiet" (Schulz, 2017, on Article 22 GDPR, point 2; see also section B.VII.2 above) about computer-made decisions seems appropriate.

The consequence of the lack of a "recognisable protective strategy" (Veil, 2018, on Article 22 GDPR, point 4) in the provision is that the scope and interpretation of the 'decision' criterion are uncertain. The uncertainty is

compounded by the fact that, according to recital 71 of the General Data Protection Regulation, the decisions covered by Article 22 GDPR also include ‘measures’ – a term which, compared with ‘decisions’, has considerably stronger connotations of practical (as opposed to legal) implications for the data subject (Veil, 2018, on Article 22 GDPR, point 58; see, however, Abel, 2018, who sees in the expression a relic of earlier regulatory ambitions of the Commission and therefore warns against overinterpretation). Accordingly, the scope of Article 22 GDPR certainly cannot be reduced, for example, to the submission of an electronic declaration of intent (for a detailed treatment of these, see Wiebe, 2002) or to other alterations of the legal position by the data subject resulting directly from a data-processing operation. Nor can the applicability of the provision be determined on the basis of a purely formal appraisal in which the crucial point is whether a person communicates the decision to the data subject or appears to the data subject to be the author of the decision. This appraisal would limit the scope of Article 22 GDPR in an entirely inappropriate way, for the data subject is in no less need of protection for having been personally notified of the result of a data-processing operation rather than, for example, through a computer interface.

Legal commentaries on Article 22 GDPR (and its forerunner provisions in Article 15 of the Data Protection Directive and section 6a of the old version of the Federal Data Protection Act – see Bygrave, 2001) offer only scant guidance for the identification of those mechanical operations which, as ‘decisions’, are to be subject to the legal consequences of this provision. They resort to explaining the definitional element by means of illustrative strategies, describing cases in which Article 22 GDPR is unequivocally applicable. This enables legal practitioners to argue by analogy and achieves a certain degree of predictability in the application of the law. It is emphasised that the provision of assistance by decision support systems is the very thing that is not meant to be covered by the provision. In other words, whenever there is a human decision-maker at the end of the chain, Article 22 GDPR is not applicable. In view of the complex and complementary interaction between machines and humans, this merely flags up the problem rather than offering a resolution criterion.

Numerous mechanical actions (on technical systems and authorship, see Schulz-Schaeffer, 2007) are embedded in complex ‘choice architecture’ (Thaler and Sunstein, 2008, pp. 89ff.; Sunstein and Reisch, 2014, p. 157), in which both human decision-makers and machines are involved. Machines prioritise, sort and classify and so channel the attention of human decision-makers (Wu, 2016, and Mengden, 2018). They preshape the spaces in which human autonomy of action operates and prestructure human decision-making processes. People do not normally take decisions in a vacuum, as it were, but weave another thread into a complex social fabric with their decision.

A robust interpretation of the conditions in which the defined ‘decision’ takes place must consider the relevant “socio-technical system” (Ropohl, 2009, pp. 58–59) in its entirety. Accordingly, a ‘decision’ may be the more plausibly assumed to be open to the application of Article 22 GDPR, firstly, the less it may be expected that a person will amend the result of the automated data-processing operation or deem it irrelevant and, secondly, the more socially significant the result of the data-processing operation is. Only then, in fact, does it make sense for the legislature to have given data subjects a right to obtain human intervention on the part of the data controller (Article 22(3) GDPR). If, on the other hand, the result of the data-processing operation were only one consideration that fed into a human decision taken on a broad factual basis, the enshrined entitlement to a “human in the loop” (Dreyer and Schulz, 2018, p. 36) would contribute nothing to the protection of data subjects.

2.3 When does a ‘decision’ fall within the scope of Article 22 GDPR?

To determine whether a given decision falls within the scope of Article 22 GDPR, it is necessary to examine the decision-making channels and routines of the data processing organisation in their entirety. It is necessary to address the question of ways in which human decision-makers incorporate the results of data processing operations into their own subsequent decisions. If the processes of “producing decisions” (Hoffmann-Riem, 2016, pp. 98ff.) attain a certain level of complexity, arguments for and against the applicability of Article 22 GDPR will become identifiable.

It should be seen as an initial indicator that an automated decision lies outside the scope of Article 22 GDPR if the baseline data on which a computer has worked are also accessible to people. The result of the data processing operation which is presented to the human decision-maker is then not a non-contextual and opaque expert judgment (on the danger of decontextualizing scores, see section B.I.5 above) but more of a ‘second opinion’. If a human decision-maker sometimes actually steps out of the pre-shaped decision-making space and overrules the automated opinion, that is a solid piece of evidence against the presence of an automated individual decision within the meaning of Article 22 GDPR. If a human decision-maker is authorised within his organisation to form his own judgement on the basis of nothing more explicit than a general appraisal of all circumstances or to take account of atypical features of individual cases, that is also an indicator that Article 22 GDPR does not apply. The situation must be assessed differently if a human decision-maker diverges from an automated preliminary decision in compliance with a particular rule. In this case there is a division of labour, with the machine and the human agent cooperating on various aspects of the final decision, so that the human decision cannot be cited against the applicability of Article 22 GDPR. It would have to be regarded as indicating an automated decision within the meaning of Article 22 GDPR if divergences from the automated verdict, though possible, carried considerable liability risks. Particularly in the medical field, this could make Article 22 GDPR applicable to a wide range of decisions.

2.4 Potential for the regulation of scoring by Article 22 GDPR

Measured against these yardsticks, scoring will often be brought within the scope of Article 22 GDPR but not as defined. A score will mostly be brought within that scope because, once it has been calculated, it will normally stick like a label to the scored person and be treated in subsequent decisions as an ‘objective fact’ about that person (see section B.I.5 above). This means that the empirical basis for the establishment of that score will no longer be subject to human appraisal – once it has been wrapped up, the score package will not be undone again. By contrast, although a particular score may, at first sight, carry weight for a human decision-maker, the latter may quite frequently access the data underlying the data subject’s score and form his or her own judgment, in which case the score cannot be brought within the scope of Article 22 GDPR.

3. Scoring of probability values (section 31 of the Federal Data Protection Act)

Section 31 of the Federal Data Protection Act sets out the requirements for “the use of a probability value for certain future action” by a natural person “for the purpose of deciding on the creation, execution or termination of a contractual relationship” with that person and defines the use of this value as “scoring”. The object of legislative regulatory activity, then, is not the prediction process as such but the handling of scores in legal transactions.⁸⁵

⁸⁵ This is new. Section 28b of the old version of the Federal Data Protection Act could easily be interpreted to mean that the prediction process itself was to be defined as scoring.

Section 31 of the Federal Data Protection Act Protection of commercial transactions in the case of scoring and credit reports

(1) For the purpose of deciding on the creation, execution or termination of a contractual relationship with a natural person, the use of a probability value for particular future behaviour on the part of this person (scoring) shall be permitted only if

1. the provisions of data protection law have been followed;
2. the data used to calculate the probability value are demonstrably essential for calculating the probability of the behaviour on the basis of a scientifically recognised mathematical-statistical procedure;
3. other data in addition to address data are used to calculate the probability value; and,
4. if address data are used, the data subject was notified of the planned use of these data before the probability value was calculated; this notification shall be documented.

(2) The use of a probability value calculated by credit reporting agencies to determine a natural person's ability and willingness to pay shall be permitted in cases where information on claims is included only as far as the conditions of subsection (1) are met and only claims concerning a performance owed which has not been rendered on time are considered

1. which have been established by a final decision or a decision declared enforceable for the time being, or if an executory title has been issued under Section 794 of the Code of Civil Procedure,
2. which have been established under Section 178 of the Insolvency Act and have not been disputed by the debtor at the verification meeting,
3. which the debtor has explicitly acknowledged,
4. for which
 - a) the debtor has received at least two written reminders after the due date of the claim,
 - b) at least four weeks have elapsed since the first reminder,
 - c) the debtor was previously informed, at least in the first reminder, of possible consideration by a credit reporting agency and
 - d) the debtor has not disputed the claim, or
5. the contractual relationship on which the claim is based can be terminated without prior notice for payment in arrears and the debtor has been informed of possible consideration by a credit reporting agency.

The lawfulness of processing, including the calculation of probability values, other data relevant for credit reports pursuant to general data protection law shall remain unaffected.

Whether it lay within the powers of the German legislature to supplement the General Data Protection Regulation with a special national provision relating specifically to scoring is a source of controversy among legal scholars (see, for example, Buchner, 2018, on section 31 of the Federal Data Protection Act, point 4). The issue could be the subject of a decision on the part of the Court of Justice of the European Union if a scorer subject to section 31 of the Act did not feel bound by the restrictions imposed by that provision. In particular, the restriction of the scope for geo-scoring (section 31(1)(3) of the Act) could offer scorers an economic incentive to test the compatibility of section 31 of the Act with EU law in court.

3.1 Scoring within the meaning of this report and of section 31 of the Federal Data Protection Act

The definition of the area requiring regulation that was the focus of legislative intervention in the form of section 31 of the Federal Data Protection Act differs in two respects from that of scoring as understood in this report: firstly, the report examines the scoring process as a whole without defining it from the outset by association with a particular social, economic or legal purpose; secondly, the report also considers scoring processes in which the score does not express the probability of particular future behaviour but is intended to influence the scored person's behaviour.

3.2 The subject matter of section 31 of the Federal Data Protection Act

Section 31 of the Federal Data Protection Act encompasses those scoring processes in which the calculated score directly represents a verdict on the probability of the future occurrence of a particular mode of behaviour. The provision uses the term 'probability value' (*Wahrscheinlichkeitswert*), which is not commonly encountered in scholarly literature on statistics, rather than simply referring to 'probability' (*Wahrscheinlichkeit*). This does not imply any substantive differentiation. The divergence from the terminology of stochastic processes, as exemplified by Christoph and Hackel (2002), is more likely due to the fact that the term 'probability' was already, to some extent, reserved for use in legal contexts. In other provisions it

is used to designate not precise values but qualitative judgements of a non-numerical nature on causal chains (see, for example, in the realm of social compensation section 1(3) of the Federal War Victims Pension Act (*Bundesversorgungsgesetz*) and section 81(6) of the Military Pensions Act (*Soldatenversorgungsgesetz*), in the law governing the police section 56(1)(2) of the Federal Criminal Police Office Act (*Bundeskriminalamtgesetz*) and in the law on damages section 252 of the Federal Civil Code (*Bürgerliches Gesetzbuch*)). By using the term 'probability value', in section 31 of the Federal Data Protection Act, the legislator actually seeks to adopt the terminological categories of stochastic processes rather than to disown them.

The scoring processes covered by section 31 of the Act express probability in the form of a figure, for example $p = 0.96$ or $p = 96\%$ or in the form of portrayals that are unequivocally translatable into probabilities or particular grades of probability. It makes no difference, for instance, whether a scorer describes the result of the scoring process as "gold status" or the like, as long as it is clear that this label is assigned to a form of future behaviour with a probability rating of 99% or above. Section 31 of the Act does not cover scores which are designed to influence behaviour, such as health scores or driver scores, but which also contain a predictive element (on this type of score, see subsection 3.3 below).

The probability value calculated by scoring practices within the meaning of section 31 of the Federal Data Protection Act must relate to particular future behaviour (*Verhalten*) of a natural person. This means that the 'event' (in the stochastic sense – see Christoph and Hackel, 2002) the probability of which is calculated in the scoring process must be legally definable as *Verhalten*. The legislator sought to limit the scope of section 31 of the Act "to self-determined action, so as to exclude events attributable to *force majeure* or to outside influence, such as lightning strikes, theft or the onset of illness. The procedures for pricing products such as life or health insurance policies or insurance against vehicle theft therefore do not constitute scoring within the meaning of section 28b [and now section 31 in the new version of the Federal Data Protection Act]" (Bundestag printed paper 16/10529, p. 16).

The self-determined action that section 31 of the Federal Data Protection Act is supposed to cover, however, is often no more than a predictable response to events that the scored person experiences. There is a close connection between falling victim to theft and making a theft-insurance claim or between becoming ill and claiming for medical costs. The focus on the definitional element ‘behaviour’ does not seem to be a very good way to restrict the scope of section 31 of the Act effectively and predictably. Its selectivity will grow increasingly fallible as scoring processes become more efficient.

The difficulties involved in restricting the scope of section 31 of the Act to behavioural predictions are already evident at the core of the provision, in credit scoring. The genesis of the scoring clause, section 28b, in the old version of the Federal Data Protection Act had everything to do with regulation of the activity of credit reference agencies. The target variable expressed in a credit score, however, is not in any way a form of behaviour on the part of the debtor but the impersonal occurrence of a negative attribute (see section C.III.3 above). Even in the wording of section 31 the vagueness of the restriction to the “self-determined action” (Bundestag printed paper 16/10529, p. 16) of the scored person is discernible. The second paragraph regulates the calculation of probability values for the scored person’s ability to pay. In some circumstances, however, this will be no more a result of the self-determined action to which the legislative explanatory memorandum refers than the onset of an illness, which the legislature certainly intended to exclude from the scope of section 31 of the Federal Data Protection Act.

3.3 Predictive scoring beyond the scope of Article 31 of the Federal Data Protection Act

To whatever extent ‘events’ are described as ‘behaviour’ in stochastic terminology (see subsection 3.2 above), section 31 of the Federal Data Protection Act does not exhaust the field of predictive scoring. Even scores that do not represent a precise verdict on probability can serve as predictors of future modes of behaviour, for it is fair to assume that the modes of behaviour of

scored persons which have contributed in the past to their scores will continue to be displayed by these persons in the future.⁸⁶ Take, for example, a health score, in which points are awarded for a motley assortment of healthy forms of behaviour, such as taking exercise and abstaining from alcohol. A high score in this case certainly warrants the generalising prediction that the modes of behaviour underlying the calculated score will be maintained at a similar level in the future. Against this backdrop, it is also legitimate to refer to this score as predictive. Section 31 of the Federal Data Protection Act, however, does not contain criteria for the regulation of such scores. That provision, as its wording indicates, is limited to scores that assign a probability value to a particular form of behaviour.

3.4 Regulatory intervention by section 31 of the Federal Data Protection Act

Section 31 of the Federal Data Protection Act addresses scoring as an issue relating solely to data privacy law. This means that data-processing operations with no input of personal data fall outside the scope of the provision altogether (section 1(1) of the Act; see also the heading of Part 2, chapter 1, of the Act: “Legal basis for processing personal data”). The fact that regulatory intervention by section 31 of the Act is specifically confined to data privacy means that quality and fairness requirements for scoring processes are only indirectly addressed by the legislation, namely as requirements for the processing of data used in the calculation of scores.

It is not self-evident that such regulatory intervention in the realm of scoring should focus specifically on data privacy. This intervention takes on sharper contours against a contrasting backdrop, namely the law governing the capital adequacy of credit institutions. A credit institution with insufficient equity is not only endangering its own existence as a business but may also have serious consequences for the stability and viability of the financial market. That is why capital adequacy of credit institutions is a key component of financial regulation. The extent to which the lending activities of credit institutions must be underpinned with equity

⁸⁶ Even the supposedly precise probability-generating scoring practices under section 31 of the Federal Data Protection Act are based on this assumption; there is nothing baffling about this, for it goes without saying that every model needs to draw on past data for its statements about the future.

depends in part on the risk of default on the loans it has issued. A loan with a low risk of default requires a lower capital adequacy ratio than a high-risk loan. This is why rating systems, in other words the models that are used to measure the appropriate capital adequacy ratio, are subject to intensive regulation (Paraschiakos, 2017, and Wundenberg, 2012).

The rating systems that are used to measure risk must meet legally defined quality requirements. The legislative foundations of the law governing the regulation of capital adequacy – see especially the Credit Institutions Directive⁸⁷ and the Capital Requirements Regulation⁸⁸ – are specialised instruments of considerable complexity.

The Banking Act (*Gesetz über das Kreditwesen*), in section 10(2), contains a provision that is similar to section 31 of the Federal Data Protection Act, right down to elements of the wording. The provision allows credit institutions to collect and use personal data in order to develop and operate rating systems. In this respect it does not differ conceptually from Article 31 of the Federal Data Protection Act, which gives scorers comparable permission for their scoring systems. One significant difference, however, is that the comprehensive quality-assurance legislation for rating systems which was outlined above exists alongside section 10(2) of the Banking Act. Section 10(2) of that Act gives privacy clearance for the pursuit of standards set by other legal provisions. In the realm of scoring, however, there are no such quality requirements stemming from other legislation.⁸⁹ Should the economic or social relevance of scoring be deemed comparable to that of the capital adequacy of credit institutions, it would be logical to supplement section 31 of the Federal Data Protection Act with a pendant in the form of provisions outside the sphere of data privacy law designed to guarantee the quality of scoring. These

provisions would not be subject to any of the objections that are raised regarding the compatibility of section 31 of the Federal Data Protection Act with EU law (see subsection 3.1 above).

3.5 The structure of section 31 of the Federal Data Protection Act

In paragraph 1, section 31 of the Federal Data Protection Act contains rules for the use of every kind of predictive score.⁹⁰ Paragraph 2 contains supplementary provisions on credit scoring, that is to say on the calculation of the probability “of a natural person’s ability and willingness to pay”.

The provision begins by establishing the primacy of the general provisions of data protection law. Every scoring process must, of course, comply with the general provisions of data protection law (section 31(1)(1)). In particular, personal data on which the calculation of the score is based must be collected in accordance with the first sentence of Article 6(1) GDPR, and the scoring process must be consistent with the principles set out in Article 5(1) GDPR. This means that a mere reading of section 31 of the Act conveys an extremely incomplete image of the protection of consumers in the context of scoring. The far-reaching safeguards enshrined in general data privacy law which do not specifically refer to scoring must also be applied to scoring operations, such as the fundamental requirement for the data subject to give consent to the processing of his or her personal data for the purpose of scoring (Article 6(1)(a) GDPR)).

In addition, section 31 of the Federal Data Protection Act lays down two types of scoring-specific requirement that must be met when scores are calculated. The provision formulates normative requirements for scoring on the one hand and quality requirements on the other.

87 Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC

88 Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012

89 The quality of scoring services is addressed very indirectly by commercial law. Section 38 of the Trade Regulation Code (*Gewerbeordnung*), on trading activities requiring supervision, provides in paragraph 1(2) for a general routine verification of the reliability of traders (for more details, see Schönleiter, 2009, on section 38 of the Trade Regulation Code, points 4 and 19ff.), thereby tightening the normal system of official trade supervision defined in sections 1, 14 and 35 of the Code. On the relationship between the subject matter of section 30(1)(2) of the Trade Regulation Code on the one hand and of section 31 of the Federal Data Protection Act on the other, see Overbeck, 2016, pp. 11–12.

90 In the legislative procedure for the old version of the Federal Data Protection Act, the Bundesrat called for section 28b in its entirety to be limited to loan agreements in which there was a risk of default (Bundesrat printed paper 548/1/08, pp. 11–12); the Federal Government argued successfully against such a restriction of the scope of section 28b on the grounds that there was no “objective justification” for it (Bundestag printed paper 16/10581, p. 1); see Domurath and Neubeck, 2018.

Section 31 of the Act addresses genuinely normative concerns in that it answers the question as to which data may be included in the calculation of the score. For example, a credit score must not be based on the mere fact that the debtor has not settled an outstanding debt. On the contrary, particular circumstances, such as a payment order issued by a civil court or the debtor's explicit acknowledgement of the claim, must also obtain before non-settlement of a debt can serve as a scoring criterion. Moreover, certain data, namely address data, which the law regards as particularly sensitive, may be used for the calculation of scores in conjunction with other data but must not be used as the sole scoring criterion (section 31(1)(3)). Certain data may not be used unless specified procedural and formal requirements have been met. This applies once again to address data (section 31(1)(4)) and also – in the case of credit scoring – to data on overdue debt settlement (section 31(2)(1) to (2)(5)). Data that have not been collected in accordance with the applicable provisions on data acquisition may not be used for scoring purposes, regardless of whether they increase the predictive performance of a scoring process. The purpose of all these requirements, then, is not only that they should be instrumental in ensuring the highest possible scoring quality; on the contrary, they also serve the paramount purpose of ensuring fairness. These requirements thus reflect ideals concerning socially appropriate forms of scoring.

Secondly, section 31 of the Act sets standards for the quality of the process of calculating scores. Here the provision answers the following question: what are the quality requirements that predictive scoring processes must meet? The provision that sets out the basic definition of these quality requirements is section 31(1)(2) of the Act, which stipulates that “the data used to calculate the probability value” must be “demonstrably essential for calculating the probability of the action on the basis of a scientifically recognised mathematical-statistical procedure”. This provision thus applies a standard of empirically scrutable rationality to predictive scoring processes.

3.6 Potential for the regulation of scoring by section 31 of the Federal Data Protection Act

Section 31 of the Federal Data Protection Act has the potential to serve as a blueprint for an appropriate overarching regulation of scoring.⁹¹ The same is less true of the normative requirements relating to admissible data inputs. Here the provision, even its generally worded paragraph 1, is recognisably tailored to the regulation of credit scoring. The prescribed limitations on the baseline data are driven by the focus on that area of economic activity. A general basis for the regulation of scoring is not discernible in those parts of section 31. The same cannot be said of the requirements relating to scoring quality, namely that data be essential and that calculations be scientifically based. In this case the current legislation contains the nucleus of a general ‘product safety law’ for scoring processes. This would entail the requirements set out in section 31(1)(2) of the Act being developed further than they have been to date (see section E.III.1 below).

⁹¹ Going beyond the limited sphere of scoring processes, the Advisory Council for Consumer Affairs examined the regulatory potential of section 28b of the Federal Data Protection Act (section 31 in the new version of the Act) in its report *Consumer Rights 2.0* and found that the provision represented a useful starting point for regulating self-learning algorithms (SVRV, 2016, p. 6, option II.1, and, for an explanation of its finding, pp. 60–61).

II. Rules for specific areas of activity

Besides the requirements of Article 22 GDPR and of section 31 of the Federal Data Protection Act relating to data processing in general, there are also rules that apply to specific areas. They address particular practical problems that scoring raises in individual economic sectors or in everyday transactions with legal implications. These rules, however, were not created in a conscious effort to regulate facets or specific manifestations of scoring. Accordingly, they do not fit into a system of scoring regulation but necessarily create a fragmented panorama. The following remarks, therefore, are not exhaustive in the sense that there could be no other legal yardsticks against which the “assignment of a numerical value to a person for the purpose of predicting or guiding that person’s behaviour” could be productively measured. We have excluded, for example, the dimension of scoring that genuinely relates to human rights and fundamental freedoms, even though scoring is liable to interfere significantly with the privacy rights of scored individuals. There is clearly a need for research in this field (see also *Gesellschaft für Informatik*, 2018).

(3) Subsections (1) and (2) above and sections 308 and 309 shall apply only to provisions in standard business terms on the basis of which arrangements derogating from legal provisions, or arrangements supplementing those legal provisions, are agreed. Other provisions may be ineffective under the second sentence of subsection (1) above, taken in conjunction with the first sentence of subsection (1) above.

A scoring system may be part of a contract between a scorer and a scored person. The calculated score has an influence on rights and obligations within a contractual relationship that has been dynamised by scoring. An illustrative example is a score that determines a level of payment, such as an insurance premium determined by the recorded behaviour of the policyholder which is translated into a score.

Like all contents of contracts, such scoring clauses – in the legally defined general circumstances (see sections 305 and 301 of the German Civil Code (*Bürgerliches Gesetzbuch*)) – must meet the requirements of the law governing standard business terms. The graduated test regime for the inclusion of clauses in a contract, their interpretation and the admissibility of their contents contains instruments which, at least in some respects, may be able to provide protection against unfair scoring. For this reason, the law governing standard business terms is a key point of reference for any discussion of the admissibility and limits of scoring clauses (Rudkowski, 2017, Brömmelmeyer, 2017, and Klimke, 2015).

Protection against the involvement of customers in scoring systems on a contractual basis against their will is provided by the obligations on users of general business terms to refer the other party explicitly to the general business terms and to give that party the opportunity to take notice of their contents (section 305(2)(1) and (2) of the German Civil Code). If the clauses on which the scoring system is based are deemed to be “surprising”, they do not form part of the contract (section 305c(1) of the Code). The practical significance of this safeguard must be considered minimal at the present time. Firstly, companies use scoring systems today as a means of interesting prospective customers in their products and ensuring customer retention (see chapter C.III above). Scoring systems, moreover, are built on a technical in-

1. The law governing standard business terms

i

Section 307 of the German Civil Code Test of reasonableness of contents

(1) Provisions in standard business terms shall be ineffective if, contrary to the requirement of good faith, they unreasonably disadvantage the other party to the contract with the user. An unreasonable disadvantage also arises from the provision not being clear and comprehensible.

(2) (...)

frastructure that would be virtually impossible to establish without the cooperation of customers. A ‘surprising’ incorporation of scoring systems into contracts therefore seems improbable.

The second sentence of section 307(1) of the German Civil Code contains an effective instrument for the regulation of scoring in terms of transparency. The provision prescribes a test of reasonableness for standard business terms, based on the question whether those terms unreasonably disadvantage the other party. Such unreasonable disadvantage may even arise from a provision not being “clear and comprehensible” (for a detailed examination of the dimensions of this requirement, see Micklitz, 2014). The provision, unlike other tests relating to the content of standard business terms, also covers what is known as the uncontrolled area, namely the main promise of performance and the price clauses (Wurmnest, 2016, on section 307 of the German Civil Code, points 12ff. and 16ff.). In the context of insurance, the second sentence of section 307(1) of the Civil Code, taken in conjunction with the second sentence of section 307(3), thus permits a review, for example, of a contractual clause establishing the principle that the premium, or a discount granted on the premium, is variable (Rudkowski, 2017).

Finally, the law governing standard business terms, through the first sentence of section 307(1) and section 307(2) of the Civil Code, as well as through sections 308 and 309, makes it possible to review contractual clauses for balanced content. This possibility covers, for instance, clauses in insurance policies that define the rules relating to the adjustment of premiums; these do not fall into the uncontrolled area defined by the first sentence of section 307(3) of the Civil Code (Rudkowski, 2017). The first sentence of section 307(1) requires a balance to be struck between the interests of the user with those of the other party (Wurmnest, 2016, on section 307 of the German Civil Code, point 33). In this balancing process, individual concerns of the scored person relating specifically to scoring may be taken into consideration. For example, it may be assumed that relatively little weight will be assigned to an interest in the implementation of scoring processes that are open to objection on quality grounds, whereas the breadth of the collected baseline data used in the calculation of a score may be a significant interest. Restriction of the

general right of privacy of the scored person is all the more significant the wider the range of his or her collected personal data (cf. Rudkowski, 2017).

2. The law governing insurance contracts and insurance supervision

An insurance policy is a comparatively heavily regulated ‘legal product’ (Dreher, 1991), and the insurance industry is a tightly supervised sector, the structure of which is determined to a considerable extent by EU requirements (Dreher, 2018, introduction, provides an overview). The importance – often considerable and sometimes even existential – of insurance policies to insured persons and the systemic importance of insurance companies to the stability and efficiency of financial markets legitimise the principle of a stringent legal regime for insurance activities.

The question of the legal admissibility of scoring processes in insurance must be answered cautiously. The answer depends on the type of insurance under consideration, because there are significant differences between the legal provisions governing the various forms of insurance.

It must also be borne in mind that scoring methods within a branch of the insurance industry can be integrated into the contractual relationship in very different ways (cf. Deutsche Aktuarvereinigung e.V., 2017). While the dynamic adjustment of premiums on the basis of scores is the most illustrative example, it is not the most commonly practised method at the present time (see section C.III.1 above).

The adjustment of premium discounts, the calculation of policyholders’ dividends or of bonuses and rebates and the determination of benefits for redemption from third parties are other areas in which scoring processes may be used. In particular, the individualisation of the content of insurance policies when they are taken out may be interpreted as a form of scoring; the considera-

tion of factors such as the age of the policyholder has a long tradition. A blanket assessment of the admissibility of scoring in insurance law is therefore precluded and would not begin to do justice to the complexity of the insurance industry (Bitter and Uphues, 2017).

General legal rules governing the supervision of the insurance industry are not fundamental obstacles to the introduction of scoring elements in insurance. Insurers can adapt their business activities accordingly. Leasing of the telematics components that are required for scoring purposes cannot, in principle, be regarded as alien to insurance activities and hence classifiable under section 15 of the Insurance Industry Supervision Act as inadmissible non-insurance business (Klimke, 2015). The prohibition of special benefits under section 48b of the Insurance Industry Supervision Act must be observed, but it is not categorically incompatible with scoring elements in an insurance policy. The special data-privacy rule in section 213 of the Insurance Contract Act (*Versicherungsvertragsgesetz*) enumerates the sources from which an insurer may acquire personal health-related data about an insured person. These sources are doctors, hospitals and other health institutions, care homes and nursing staff, other insurers of persons and statutory health insurance funds as well as social insurance funds for occupational accidents and public authorities. Providers of wearable devices or health apps for smartphones (Adam and Micklitz, 2016, have already addressed these) are not listed there. That provision, however, is unlikely to be a serious obstacle to the establishment of scoring elements in health insurance, because first of all the provision is subject to the disposition of the parties (Rixecker, 2016, on section 213 of the Insurance Contract Act, point 28) and, secondly, it can be plausibly argued that, where an insurer accesses data generated by wearable devices or apps, the data have not been collected from third parties but from the insured person.

This report focuses sharply on motor and health insurance as potential arenas of scoring activity.⁹² This approach brings the two extremes of the regulatory spectrum into view. In the case of motor insurance, current legislation relating specifically to insurance places

scarcely any obstacles in the way of the integration of scoring elements into motor policies. Private health insurance, by contrast, because of its vital importance to insured persons (Schüffner and Franck, 2018, on section 47 of the Insurance Industry Supervision Act, point 118) and its socio-political relevance, is enclosed in a tight regulatory straitjacket. In this area the law as it stands leaves little scope for the use of scoring processes during the lifetime of a current policy.

Motor insurance is very largely open to the introduction of scoring processes. The guiding principle of loss indemnification based on collective solidarity is alien to this type of insurance. The terms of insurance are determined by the principle of equivalence, whereby the premium is set in relation to the assumed risk. A *de facto* ‘solidarity effect’ may be discerned in the fact that, within a period of cover, some people from the pool of policyholders assigned to a particular risk group will incur losses, while others will not. This, however, amounts to “compensating for random fluctuations in claims experience and over time and not compensating for systematic differences in the gravity of individual risks” (Bitter and Uphues, 2017, pp. 3–4). This solidarity effect is intensified by the fact that the technical means of actuarial ‘fragmentation’ (Looschelders, 2015) of the body of policyholders into individual risk groups are not yet fully developed. The result is a levelling process between policyholders with a high probability of loss and those whose probability of loss is lower, all of whom are lumped together in a single risk group.

In these circumstances it comes as no surprise that no structural legal obstacles to the introduction of telematics-based pay-as-you-drive tariffs are identified (Klimke, 2015; Schumann, 2017; Koch, 2017, commentary on the General Conditions for Motor Insurance of 2015, points 9ff.). One issue could, however, be the compatibility of telematics-based tariffs with the provisions of insurance contract law on aggravation of risk (sections 23ff. of the Insurance Contract Act); for more details, see Lüttringhaus, 2018. There are no objections in principle to the use of telematics data in legal proceedings relating to the occurrence of an insured event (Klimke, 2015).

92 On scoring, albeit without the use of that term, in life and occupational disability insurance, see Brömmelmeyer, 2017.

Private health insurance is a highly variegated field of diverse insurance types (full health cover and various supplementary policies – see also section 192 of the Insurance Contract Act), based on differing calculation systems (calculation of premiums as for life assurance or as for indemnity insurance – see Kalis, 2018, section 44, points 209ff.) and also performing, to varying extents, a socio-political function. The socio-political significance is obvious in the case of substitutive health insurance (section 146 of the Insurance Industry Supervision Act and section 195 of the Insurance Contract Act), which can take the place of statutory health insurance and must therefore meet special requirements. Here, more than in the other areas of insurance law, attempts to generalise are thwarted by a wide diversity of conceivable insurance products and applications for scoring methods. At the heart of the scoring debate is the conception of a future in which insurance premiums are linked to the policyholder's scored health-related behaviour. At least here the existing law on insurance supervision and insurance contracts are erecting effective barriers.

If health insurance premiums are calculated like those for life assurance, which is a prerequisite for the use of private health insurance as a substitute for statutory cover (see section 146(1) of the Insurance Industry Supervision Act), any adaptation of premiums based on scoring would come into conflict with section 203 of the Insurance Contract Act.⁹³ Paragraph 1 of that provision definitively determines (Voit, 2018, on section 203 of the Insurance Contract Act, point 6) which criteria are to be used for the calculation of the premium. Paragraph 2, taken in conjunction with the first sentence of section 208 of the same Act, lays down the conditions and the procedure for adjusting premiums. This provision is semi-mandatory (Boetius, 2017, on section 203 of the Insurance Contract Act, point 51), meaning that derogation to the detriment of the policyholder is not permitted. Neither in the determination nor in the adjustment of premiums does section 203 of the Insurance Contract Act permit the use of an individual's health-related behaviour or of a score based on that behaviour as

a basis for setting the amount of the premium,⁹⁴ for the first sentence of section 203(1) and the fourth sentence of section 203(2) of the Insurance Contract Act refer to the provisions of the Insurance Industry Supervision Act on the calculation of premiums. In this respect the provisions establish consistency between the two areas of insurance law and transform the stipulations of supervisory law into insurance contract law (Boetius, 2017, on section 203 of the Insurance Contract Act, points 3 and 42). sections 6(1) and 10(1) of the Private Health Insurance Supervision Ordinance (*Krankenversicherungsaufsichtsverordnung*) contain statutory stipulations on the calculation of premiums by private insurers. These provisions rule out consideration of the policyholder's lifestyle during the term of a policy as a calculation factor. The amount of the premium is to be based on the policyholder's age and the scope of the benefits offered by the policy (but see section 152 of the Insurance Industry Supervision Act on the specific rules governing the basic tariff, addressed in Vogt, 2018, on section 203 of the Insurance Contract Act, points 11ff.). Previous illnesses may be factored into the premium by way of risk loading (Voit, *ibid.*, point 6); otherwise there is no scope for customised premiums.

Consideration should also be given to the second sentence of section 194(1) of the Insurance Contract Act. That provision specifies that the general provisions on aggravation of risk in sections 23ff. of the said Act do not apply to health insurance. This exclusion of applicability relates not only to health policies with life assurance-type premium calculation, which is already effected in accordance with section 203 of the Insurance Contract Act (Kalis, 2017, on section 194 of the Insurance Contract Act, point 24), but also to health policies with premiums calculated in the manner of an indemnity policy. "Changes in the insured person's state of health occurring after the conclusion of the contract do not affect either the promise of performance once it has been made or the amount of the premium" (Kalis, *ibid.*). Any subsequent aggravation of risk is borne by the insurer and not the insured (see the Federal Administrative

⁹³ The applicability of section 203 of the Insurance Contract Act, however, does not depend on the private health insurance in question actually meeting the conditions set out in section 146 of the Insurance Industry Supervision Act; see Voit, 2018, on section 203 of the Insurance Contract Act, point 5.

⁹⁴ The second sentence of section 203(1) of the Insurance Contract Act, however, provides scope for individualisation, stipulating that, "Other than with contracts in the basic tariff in accordance with section 152 of the Insurance Industry Supervision Act, the insurer may agree an appropriate risk premium or release from obligation to effect payment, taking account of an aggravation of the risk insured".

Court judgment of 5 March 1999 – case 1 A 1/97 – and the Federal Court of Justice judgment of 9 May 2012 – case reference IV ZR 1/11).

It is also a plausible assumption that the rule set out in the first sentence of section 146(2) of the Insurance Industry Supervision Act, taken in conjunction with section 138(2) of the same Act, precludes certain score-based tariffs if they are not self-financing and are therefore indirectly subsidised by policyholders on the standard tariff (Brömmelmeyer, 2017).

Calculation of contribution rebates on the basis of scoring elements is possible in principle. This is a further development of those rules of insurance contract law that make any non-profit-related premium rebate dependent on non-recourse to insurance benefits in a previous contribution period. (Boetius, 2017, on section 203 of the Insurance Contract Act, point 322; on the more restrictive conditions governing statutory health insurance, see section 53(2) and (9) of Book V of the German Social Code). The internal consistency of a regulatory regime that is split in this way, with stringent requirements for the adjustment of premiums yet relatively generous scope for scoring in all other respects, is open to criticism. From an economic point of view, after all, contribution rebates may be deemed equivalent to an individualisation of premiums.

3. Social insurance law and statutory health insurance

The system of statutory health insurance is a self-governing public institution based on the welfare state principle. Together with the other branches of the social insurance structure, it serves to protect people against life's elementary risks and is therefore an instrument of social policy. Its structural hallmark is the principle of solidarity (Kingreen, 2003). Membership of the statutory health insurance scheme is prescribed by law for many groups of people. As a matter of principle, statutory health insurance contributions are earnings-related (German Social Code, Book V, section 241), while benefits are determined by individuals' medical needs (Butzer, 2001, and Kingreen, 2003). Neither at the start of membership nor during it is any risk-weighting of individuals' premiums permitted.

There is little room in such a system for individualised incentive systems based on scoring (see Hesse Regional Social Court judgment of 4 December 2008, case reference L 1 KR 150/08 KL, in the *Juris* database). Accordingly, there are no complex scoring processes in that domain at the present time.⁹⁵ However, the bonus programmes for health-conscious individuals for which section 65a of Book V of the German Social Code provides (Bundesversicherungsamt, 2016) constitute a system that should be described as proto-scoring. Within the system of statutory health insurance, it is an exception in particular need of legitimisation. The Federal Insurance Office takes a sceptical view of these programmes on the whole (Bundesversicherungsamt, 2018; see also section C.III.3 above).

⁹⁵ This does not say anything about the limits imposed by constitutional law on any attempt to shift the system of social insurance away from the solidarity principle towards more highly accentuated individualisation. Although the principle of the welfare state that is enshrined in Article 20(1) and the first sentence of Article 28(1) of the Basic Law requires the creation of social security systems as protection against the vicissitudes of life (Decisions of the Federal Constitutional Court, Vol. 28, p. 324, esp. pp. 348ff., Vol. 45, p. 376, esp. p. 387, and Vol. 68, p. 193, esp. p. 209; see Axer, 2000), this does not bind those systems to the structural principles of statutory health insurance, which predate the Constitution in any case.

III. Building blocks for a scoring regime

Our review of the rules relating to scoring in data privacy law and in sectoral legislation has shown that, while they do not form a legal regime governing scoring in general, they are not powerless in the face of the scoring phenomenon. The law as it stands contains instruments with which certain social challenges connected with scoring can be met.

In this chapter we intend to examine three scoring-related challenges – score quality, transparency of the scoring process and non-discrimination – and outline the extent to which current law sets standards in these domains. Since these three challenges are not confined to specific areas of economic activity, such as health insurance, or particular situations with legal implications, such as the conclusion of sales contracts, we intend to focus primarily on those rules that address the identified problems in the most general form possible.

Scoring is a data processing operation, which means that the requirements set out in the General Data Protection Regulation and in the Federal Data Protection Act are especially fruitful in achieving regulatory objectives on a broad front. When it comes to protection against discrimination, the focus turns to the General Equal Treatment Act (*Allgemeines Gleichbehandlungsgesetz*). Although its scope is substantively limited (sections 2 and 19) and many of its legislative details reveal the compromise character of the protection against discrimination that has ultimately been achieved, it not only contains legal definitions of direct and indirect discrimination (section 3(1) and (2)) but is also suitable as a model with the aid of which the general review programme for the identification of unlawful cases of unequal treatment can be explained.

1. Regulating the ‘how’ of scoring versus regulating the ‘whether’

The challenges referred to above of mathematical and statistical quality, transparency and non-discrimination relate to the phenomenon of scoring **when it occurs**. We do not, on the other hand, explore in depth the preliminary question of the legal relationships within which scoring is even permissible. The question regarding the ample scope for scoring processes can only be answered discretely for specific areas of people’s lives and of economic activity.

The regulatory regimes for scoring in the insurance industry which are outlined above illustrate that provisions in particular areas of activity may be obstacles to the application of scoring methods. Such an obstacle to scoring may result from the fact that the criteria governing the way in which private individuals may act and take decisions are set in stone, one example being the requirements outlined above for the calculation of premiums and premium adjustments in the realm of private health insurance. The law, however, may also prevent scoring by prohibiting the use of the very knowledge that is obtained from scoring processes. In the areas of economic activity that are examined in this report, there are no legislative examples of this type of regulatory model. It may therefore be expected that, with advances in the technological scope for the use of scoring methods and their spread within the economy, the legislature or the judicature – possibly by affirming the horizontal effects of the general right of privacy – will clamp down specifically on the use of scoring methods in principle in particular cases.

2. Scoring regulation and algorithm regulation

Regulation of the ‘how’ of scoring is part of the context of the regulation of algorithms, the possibilities of which are the focus of much discussion at the present time (see section B.I.4 above). Accordingly, in its report entitled *Consumer Rights 2.0*, the SVRV stated that the scoring provision in the Federal Data Protection Act⁹⁶ contained a legislative starting point for the regulation of algorithms (SVRV, 2016; see also Härting, 2015). This is the basis for the following reflections.

The special case of scoring regulation may serve to illustrate how the law can formulate and enforce quality requirements, as exemplified here by quality assurance, and ethical requirements, as exemplified by the prohibition of discrimination, for algorithms. The study conducted by the Specialist Group on Legal Informatics of the German Informatics Society entitled *Technische und rechtliche Betrachtungen algorithmischer Entscheidungsfindung* (‘Technical and legal reflections on algorithmic decision-making’) – is therefore one of the main sources on which the following remarks are based (cf. also Gigerenzer, Wagner and Müller 2018).

3. Guaranteeing a defined score quality

Numerous scoring processes serve the purpose of delivering a particular predictive service. The score is the verdict on the probability that a person will behave in a certain way in the future. This service may be rendered more or less well. Predictive scoring systems thus have a ‘quality dimension’ (on the factual situation, see section C.III.3). While the legislature is not indifferent to scoring, the requirement of scientifically recognised procedure set out

in section 31(1)(2) of the Federal Data Protection Act ensures only a minimum degree of quality. Section 31(1)(2) specifies that “the data used to calculate the probability value” must be “demonstrably essential for calculating the probability of the action on the basis of a scientifically recognised mathematical-statistical procedure”. The provision thus addresses two distinct problems,⁹⁶ both of which are rooted in the use of the formula “the data used”, because “the data used” could be intended to refer to the **general** data categories that are used in the calculation of scores, such as ‘age’, ‘number of current accounts’ or ‘address’. Or else the phrase could refer to the input variables that are used to calculate the score for a specific person (‘50’, ‘3’, ‘1 Castle Street’). There are differences between the regulatory resources which are needed to remedy defects in non-case-specific scoring formulae on the one hand and in input data relating to individual cases on the other.

3.1 Accuracy of specific input data

The concern that the data used as input variables for scoring purposes may not reflect the true factual situation (see chapter B.V above) is an aspect of the quest to guarantee high-quality data. The problem is of great practical relevance (see section C.III.3 above). Indeed the inaccuracy of personal data must be the absolutely central everyday problem in scoring-related complaints (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014, which contains reports on credit scoring from several consumer advice centres), to go by the statistics from various regional consumer advice centres on the main focal points of their information activity with regard to credit scoring (see, for example, Verbraucherzentrale Bremen, 2016, Verbraucherzentrale Niedersachsen, 2015, and Verbraucherzentrale Nordrhein-Westfalen, 2018).

Ensuring that only correct personal data are processed, however, relates not only to scoring but essentially to every data-processing operation. It therefore seems to be an excessively restricted and hence inappropriate approach to the problem to address it largely through the clause in section 31(1)(2) of the Federal Data Protection

⁹⁶ In addition, the requirement of demonstrability is significant in its own right, since it establishes a documentation requirement for scorers; see Bundestag printed paper 16/10529, p. 16, and Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein and GP Forschungsgruppe, 2014. This aspect is not addressed in the following paragraphs.

Act which prohibits the processing of inaccurate data (and hence also non-essential data within the meaning of section 31(1)(2) of the Federal Data Protection Act – see Overbeck, 2016). In the longer term there will be a need for a body of law designed to ensure the quality of stored data. A normative mooring for such a legal regime already exists today in the principle of data accuracy enunciated in Article 5(1)(d) GDPR (see Pötters, 2018, on Article 5 GDPR, point 24). The contours of this area of the law and of the obligations that controllers have to fulfil with regard to the accuracy of the data they process, however, have scarcely been developed at all to date (Hoeren, 2016). One legally simple way of overcoming this problem certainly lies in the rights of data subjects to information and rectification (first sentence of Article 16 GDPR; for more details see Domurath and Neubeck, 2018). In this respect, however, data privacy law suffers from a considerable mobilisation deficit (Härting, 2015; Spindler, Thorun and Wittmann, 2017).

3.2 Scientific basis of scoring processes

Section 31(1)(2) of the Federal Data Protection Act prescribes that scoring processes must meet certain scientific standards (see section B.IV.1 above). With this provision, the legislature excludes at least the use for scoring purposes of data that cannot contribute anything to the predictive performance of a scoring process (Domurath and Neubeck, 2018). Where it is not even possible to demonstrate a correlation between a particular type of data and the event whose probability is to be predicted, the use of that type of data would be contrary to section 31(1)(2) of the Federal Data Protection Act.

Attempts are sometimes made to frame requirements for the instrumentality of the data that are used which go beyond proof of correlation. Formulating these requirements in such a way that they are usable in the practical application of the law has proved to be a difficult undertaking. This applies, for example, to the case that is sometimes made for the restriction of usable data to those that are “contractually relevant” (Domurath and Neubeck, 2018, cite examples). The types of data in question are those that influence the probability of the target behaviour in a particularly direct way (see also Buchner, 2018, on section 31 of the Federal Data Protection Act, point 8, who states that such a provision would require users “to demonstrate plausibly and verifiably

that the data used to calculate the score are of direct relevance”). It remains unclear how the catalogue of these data types could ever be reliably defined.

The ‘correlation’ requirement laid down in section 31(1)(2) of the Federal Data Protection Act implies that those who undertake predictive scoring processes must never rely, when designing the process, on statistically unverified routine plausibility checks. In this respect, scoring needs “empirical reinforcement”. This requirement is far from self-evident, for there is no general obligation on those who enjoy the fundamental right of freedom of expression enshrined in the first sentence of Article 5(1) of the Basic Law to confine themselves to rationally justified utterances, not even when they are communicating alleged facts. Seen in that light, the rationality requirement in section 31(1)(2) of the Federal Data Protection Act already looks like a thoroughly significant legislative intervention, although, given the social significance of communicated probability scores, a plausible justification can be found for it.

That a process is scientific within the meaning of section 31(1)(2) of the Federal Data Protection Act is not guaranteed solely by the fact that its predictive performance is delivered with a level of reliability appropriate to the relevant area of people’s lives. The fact is that any process which delivers a better predictive performance than the toss of a coin can be the result of proficient application of statistical methods and, as such, constitute a significant and praiseworthy scientific achievement. But it does not answer the question whether the procedure should or should not be applied in a particular area of people’s lives. Specific quality criteria are not associated with the obligation to follow a scientific procedure. In this respect the legal regime covering predictive scoring has a regulatory void, which becomes particularly striking when contrasted with something like the law governing the capital adequacy of credit institutions, which was outlined above (see subsection E.I.3.4. This does not mean that section 31(1)(2) of the Federal Data Protection Act is a toothless tiger, but it does have biting inhibitions.

4. Guaranteeing transparency and comprehensibility

The General Data Protection Regulation explicitly anointed transparency as a principle to which all processing of personal data must adhere. The third principle set out in Article 5(1)(a) GDPR is that personal data must be “processed in a transparent manner in relation to the data subject”. This principle of transparency is developed programmatically in recitals 39, 58 and 60 of the GDPR. The circuitous wording of the cited sources must not obscure the fact that the level of abstraction of the transparency principle is still considerable. Which precise duties are actually incumbent on the controller in respect of each specific data processing operation remains uncertain (see above before section E.I.1 and, for example, Roßnagel, 2018, Wachter, Mittelstadt and Floridi, 2017, and Selbst and Powles, 2017). The catalogue of obligations is fleshed out somewhat in Articles 12 to 15 GDPR.

Article 12 GDPR

Transparent information, communication and modalities for the exercise of the rights of the data subject

The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. (...)

Article 13 GDPR

Information to be provided where personal data are collected from the data subject (Article 14 is similar: Information to be provided where personal data have not been obtained from the data subject)

(...) In addition to the information referred to in Paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing:

(...)

the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

i

Article 15 GDPR

Right of access by the data subject

The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

(...)

the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

The provisions prescribe the fulfilment of extensive information obligations to the data subject (WP 29, 2018), and give the latter far reaching rights of access to information, which are also rooted in fundamental rights (second sentence of Article 8(2) of the Charter of Fundamental Rights of the European Union). But these provisions likewise leave considerable latitude for the application of the law. This has two reasons.

First of all, interests that conflict with the principle of transparency have also been recognised and must therefore be taken into account in the interpretation of the neutrally framed terminology of the transparency regime. Recital 63 makes this clear, stating that “A data subject should have the right of access to personal data which have been collected concerning him or her, and to exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing. (...) That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject.” It is recognisable that a regulation problem has been identified here but not resolved. The General Data Protection Regulation is unable to establish consensus on the issue of how the

interest of safeguarding trade secrecy and that of access to information are balanced in current data privacy law. Secondly, the General Data Protection Regulation, in what are key provisions in terms of scoring transparency, defines the catalogue of obligations

incumbent on the controller in a conspicuously uninformative manner. Article 13(2)(f), Article 14(2)(g) and Article 15(1)(h) GDPR each define information about “the logic involved” (*la logique sous-jacente; die involvierte Logik*) in automated decision-making within the meaning of Article 22 GDPR. It might be supposed that, in the disciplines in which algorithms feature, the term ‘logic’ related to an algorithm as described from a particular perspective and that the legislature had made reference to this non-legal term with a view to preparing it for reception by the legal community (examples of such processes are described in Klement, 2006, and Mathis, 2017). This supposition is wide of the mark. Mathematicians, computer scientists and software engineers have a no less vague notion than legal scholars as to what the “logic involved” in automated decision-making might be.

The lively debate (see section B.I.3 above) on the disclosure of the attributes used as input variables in Schufa credit scores and their weighting is indicative of the lack of normative guidance provided by the transparency regime of the General Data Protection Regulation. If we assume that the calculation of a Schufa score amounts to decision-making within the meaning of Article 22 GDPR, it is still a moot point which items of information on the genesis of a score are covered by the description “the logic involved” (evidence of views on the scope of the provisions can be found in Wischmeyer, 2018; for a more restrictive interpretation, see, for example, Paul and Hennemann, 2018, on Article 13 GDPR, point 31; for a broader interpretation, see, for instance, Bäcker, 2018, on Article 13 GDPR, point 54). It is sometimes assumed, by explicit reference to the Schufa judgment of the Federal Court of Justice, that the obligation to give access to information goes further than the boundaries set by the current legal position. As Florian Schmidt-Wudy writes, “With regard to the scope of the information on the “logic involved”, it remains to be seen whether the non-disclosure, approved by the Federal Court of Justice, of the scoring formula will remain tenable, for without knowledge of the scoring formula, it is scarcely possible for the

data subject to discover and correct errors in the score (...). On the other hand, unrestricted disclosure of the score may jeopardise the business model of credit reference agencies (...).

Because of the analogous application of Article 15(4) GDPR, however, and the balance it prescribes with fundamental rights and freedoms, strict secrecy of scoring formulae as approved by the Federal Court of Justice will not be maintainable if knowledge of them is essential for a data subject to be able to identify flawed calculations and have them corrected. On the contrary, it will depend on the individual case, which means that in certain cases both the scoring formula and its underlying parameters may certainly be the subject of a disclosure.” (Schmidt-Wudy, 2018, on Article 15 GDPR, point 78.3). The cautious way in which the commentator expresses his interpretation of the law, is illustrative of the strikingly weak normative guidance provided by Articles 13 to 15 GDPR (but see Heuzeroth and Seibel, 2018). The present legal position is still lagging behind the normative guidance provided by section 34 of the Federal Data Protection Act (old version), on the basis of which the Federal Court of Justice outlined the information access claim against Schufa – and that provision itself is far from unequivocal.

In the light of the above, it is no surprise that the scope of transparency requirements arising from the General Data Protection Regulation is a subject of controversy. The crystallisation point in the debate is the question whether the GDPR grants the data subject a ‘right of explanation’ of an automated individual decision. The object of this discussion, conducted on an international scale, is to build a bridge between, on the one hand, the transparency requirements of the General Data Protection Regulation and, on the other hand, the lively discussion on ways of making complex algorithmic decision-making systems comprehensible to people (see section B.I.4 above as well as *Gesellschaft für Informatik*, 2018, Selbst and Powles, 2017, Selbst and Barocas, 2018, and Wachter, Mittelstadt and Floridi, 2017).

It is certainly unmistakable that, in its transparency requirements, the General Data Protection Regulation follows on from its forerunner in EU law, the Data Protection Directive. This suggests a very cautious interpretation of the transparency requirements set out in Articles 13 to 15 GDPR (Wischmeyer, 2018). The information to be disclosed under these provisions would then be kept very general and would be confined to a merely superficial presentation of the program functions. On the other hand, this cautious circumscription of the transparency requirements in data privacy law may reflect the fact that the question how it is possible in practice to establish transparency (see section B.I.4 above) is still under discussion. At the heart of the transparency debate at the present time is not legal permissibility but technical feasibility. (see Selbst and Barocas, 2018, Burrell, 2016, and Lipton, 2016). The technical-sounding but substantively vague description of the transparency entitlement, with terms like “the logic involved”, “significance” and “envisaged consequences”, may therefore prove to be especially receptive to future developments in legal scholarship.

5. Guaranteeing non-discrimination

i

Section 1 of the General Equal Treatment Act Purpose

The purpose of this Act is to prevent or to stop discrimination on the grounds of race or ethnic origin, gender, religion or belief, disability, age or sexual orientation.

i

Section 3 of the General Equal Treatment Act Definitions

- (1) Direct discrimination shall be taken to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation on any of the grounds referred to under Section 1. (...)
- (2) Indirect discrimination shall be taken to occur where an apparently neutral provision, criterion or practice would put persons at a particular disadvantage compared with other persons on any of the grounds referred to under Section 1, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.

5.1 Discriminatory acts and discriminatory effect

It is difficult for current anti-discrimination law to accommodate the problem of discriminatory scoring in its conceptual framework (see chapter B.II above), because it typically checks whether the reasons that people or institutions give for their actions are legitimate from an anti-discrimination perspective. Whether a reason for an action is objectionable on grounds of incompatibility with anti-discrimination law may be ascertained in the following two steps:

In the first step, the question to be asked is whether the motive for the behaviour being tested for conformity with the law requires attention in the light of anti-discrimination law. This may be so because one of the grounds listed in section 1 of the General Equal Treatment Act was a determinant factor for the behaviour in question. Current anti-discrimination law. To take an example, someone refuses to conclude a contract on grounds of the other party's ethnic origin (see section 3(1) of the General Equal Treatment Act). Closer scrutiny is also called for, however, in the case of modes of behaviour with seemingly innocuous motives if those motives are particularly detrimental to any persons on account of one of the grounds listed in section 1 of the General Equal Treatment Act. For example, someone refuses to conclude a contract because of the other party's insufficient knowledge of the German language (see section 3(2) of the General Equal Treatment Act). The second step involves an examination of whether reliance on the suspect ground is justified in the given situation. At the end of this examination, it will have been established whether or not prohibited discrimination has taken place. To discriminate unlawfully, then, means to act on prohibited grounds (for a detailed treatment, see Schramm, 2013). Anti-discrimination law is 'input-focused'. Its attention is fixed on the interaction of certain decision-making criteria and their admissibility. In the realm of scoring, this method of applying the law may have unwanted results. For instance, a seller declines to do a deal with a prospective buyer because of the latter's low score. In so doing, the seller is not acting on the basis of a protected characteristic but simply of a score. This ground for refusal does not alter the fact that the sex of the prospective buyer, for instance, played a significant role in the calculation of the score. It could be argued, on the basis of that fact, that this is a case of unequal treatment requiring attention in the light of anti-discrimination law (Moos and Rothkegel, 2016, advance this argument; see also section C.III.5 above). The seller, of course, does not refuse to enter into a contract **because of** the other party's sex but because of the inadequate score. Although attempts can be made to bring such cases into the ambit of anti-discrimination law by means of rules on indirect discrimination, that will not resolve the difficulties.

Many individual variables, possibly even inestimably many, go into the calculation of a score. An audit that examined every one of the input variables and assessed its admissibility under anti-discrimination law would be practically unmanageable. Moreover, such an audit would run the risk of discovering mere spurious correlations and highlighting them as requiring justification, even though their occurrence is virtually inevitable in sufficiently large volumes of data.

For complex scoring processes, the ‘input-focused’ analysis of compatibility with anti-discrimination law must be supplemented by an ‘impact-focused’ analysis. In other words, attention should not be fixed solely on the decision-making criteria but also on the effects that decisions have. The European legislature displayed a delicate linguistic touch when providing for the possibility that motive-related anti-discrimination law would reach its limits when confronted with complex data-processing operations, and hence with scoring. In recital 71 of the General Data Protection Regulation, it does not speak of “discrimination” arising from data-processing operations but of “discriminatory effects”.

An impact-focused analysis can establish its legal bearings by reference to the fact that section 3(2) of the General Equal Treatment Act refers not only to “an apparently neutral provision” or “criterion” but also to a “practice” (*Verfahren*). By means of the term “practice”, anti-discrimination law releases itself from confinement to the scrutiny of individual motives. It opens the door to analyses which can identify even complex and scarcely penetrable “practices” (Block, 2018, on section 3 of the General Equal Treatment Act, point 69) as problematic in terms of anti-discrimination law. As far as the question of the reference point is concerned, neither national anti-discrimination law nor the underlying European directives stipulate that individual criteria must give rise to disadvantages but even permit a general overview of several provisions or entire processes (Schiek, 2007, on section 3 of the General Equal Treatment Act, point 33). Starting points for an anti-discrimination regime that does not focus primarily on motives for decisions but on the results of systems that operate in incomprehensible ways are also to be found in the case law of the Court of Justice of the European Union. German anti-discrimination law is shaped to a great extent by this case law, which has found that pay structures may be judged dis-

criminatory on grounds of gender inequality even without the need to isolate individual discriminatory factors (ECJ judgment of 27 October 1993 in case No C-127/92 – *Enderby* [EU:C:1993:859]).

An ‘impact-focused’ consideration of scoring practices relates not only to the individual variables that go into the calculation of the score but also to the effects of the scoring process. The scoring process itself is the “practice” (*Verfahren*) within the meaning of section 3(2) of the General Equal Treatment Act that must be guaranteed compatible with anti-discrimination law (Hacker, 2018).

5.2 Challenges posed by impact-focused protection against discrimination

The difficulties involved in trying to remedy the discriminatory effects of scoring on particular groups of people can be considerable. Still comparatively easy to address are those discriminatory effects that are attributable to flaws in the technical design of the scoring process. If the process produces quality disparities for various groups of people (see section B.II.3 above) and those disparities could have been avoided at no extra cost, the scoring practice is incompatible with anti-discrimination law (Hacker, 2018). Additional costs may also be imposed on the scorer if a greater degree of freedom from discrimination is thereby achievable (Hacker, 2018). However, in cases where the discriminatory effects of a scoring practice also increase its predictive power, an opportunity is provided for the scorer to justify these discriminatory effects (for more details, see Hacker, 2018).

Then there are the difficulties that arise when it comes to proving the discriminatory effects of a scoring practice (Hacker, 2018). A plaintiff who suspects discrimination will not have the comparative data to underpin his assertion (Hacker, *op. cit.*, Hildebrandt, 2015) and demonstrate the discriminatory effects of a scoring practice. And even from a bird’s eye view from which a wide panorama of data sets could be seen, it would still be hard to identify the discriminatory effects of a scoring practice, for apart from age and gender, data on the attributes that are customarily at the root of discrimination traditional are generally unavailable. “even to collect them would be problematic, because no one may be required to disclose his or her sexual orientation or religion. The establishment of ethnic origin raises a

fundamental problem, namely whether there are ‘objective’ factors at all for determining a person’s ethnicity other than his or her nationality.” (Grünberger, 2013, p. 664; cf. also Article 9 GDPR).

A simple enlargement of the material scope of the General Equal Treatment Act to include “automated decision-making practices” would not suffice to deal with the problem of discriminatory scoring. Data privacy law offers potential for a solution. The principle that personal data should be processed “fairly” (*de manière loyale; nach Treu und Glauben*) enshrined in Article 5(1)(a) GDPR is a normative anchorage in this respect (Hacker, 2018). In the case of remediable discriminatory quality differences in scoring practices, the principle of accuracy (Article 5(1)(d) GDPR) is also affected (Hacker, op. cit.). If these legal principles open data privacy law to the normative objective of protection against discrimination, rights of access to information under Article 15(1)(h) GDPR and data protection impact assessments under Article 35 GDPR will offer ways of addressing the problem of discriminatory scoring (for more details, see Hacker, 2018).

IV. Supervision

The legal requirements for scoring relate to various aspects of scoring, among the most prominent of which are transparency, quality and non-discrimination. The requirements differ, depending on who is scoring whom, for what purpose and in what way. This wide diversity of material requirements is matched by a wide array of legal implementation mechanisms. Some impression of the diversity of conceivable institutional law-enforcement arrangements can be obtained from data privacy law alone, which plays a key role in the formulation of legal requirements for scoring (see Schantz and Wolff, 2017, pp. 295ff.). Among these are legal remedies for adversely affected individual consumers, scope for class actions on the part of plaintiffs such as consumer advice centres and even rules for business organisations (Spindler, 2011, provides a comprehensive review, dealing with data protection on pp. 270ff.) as well as state supervisory mechanisms ensuring that scoring is conducted in compliance with the law.

State supervision is the key instrument for enforcement of the aforementioned quality requirements for predictive scoring prescribed by section 31 of the Federal Data Protection Act. BaFin, the Federal Financial Supervisory Authority, oversees compliance with the quality requirements governing models for the assessment of credit default risks (see section E.III.3 above). For the bonus programmes of statutory health insurance funds (see section E.II.3 above) there are legal bases that allow particularly close supervision (Ullrich, 2018, on section 65a of Book V of the German Social Code, point 7). It is possible to add to the substantive law the observance of which these bodies oversee and so to extend their respective supervisory missions.

On these grounds alone, there appears to be ample potential for sovereign supervision of scoring, because issues of confidentiality – in the sense of trade secrecy, for example – and of adverse social consequences of transparency (see section B.I.1 above) do not arise in this context, for disclosure to the state supervisory authority does not extend *ad infinitum* the circle of those who know how the relevant scoring algorithm works. The overseeing state functionaries, for their part, can be sworn to secrecy, and indeed they already are as a rule (see, for example, section 30 and section 29(2) of the Administrative Procedure Act (*Verwaltungsverfahrensgesetz*)).

Seen in this light, the problem of official supervision is one of adequate staffing and equipping of the relevant supervisory authorities. These must be enabled to conduct even complex audits of compliance with substantive law (see SVRV, 2016; also *Gesellschaft für Informatik*, 2018).



Advisory Council for Consumer Affairs

The Advisory Council for Consumer Affairs is an advisory body of the Federal Ministry of Justice and Consumer Protection (BMJV). It was set up in November 2014 by the Federal Minister of Justice and Consumer Protection.

The Advisory Council for Consumer Affairs is tasked with using research findings and drawing on the Federal Ministry of Justice and Consumer Protection's practical experience to help shape consumer policy.

The Advisory Council is independent and is based in Berlin.