Discussion Paper Ethics of Algorithms #6

Machines in the Service of Humankind

Creating AI that sustains and supports social needs

- Working paper -

BertelsmannStiftung

Machines in the Service of Humankind

Creating AI that sustains and supports social needs - Working paper -

Julia Krüger
Konrad Lischka
on behalf of the Bertelsmann Stiftung

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Managing editors
Carla Hustedt, Ralph Müller-Eiselt

Authors
Julia Krüger, Konrad Lischka

DOI 10.11586/2018061 https://doi.org/10.11586/2018061

Contents

1 Preface

Which secondary schools are children allowed to attend? Which areas do police decide to patrol most intensively? Whose tax returns are processed by humans, and whose are processed entirely by software? Which passersby in train stations are deemed suspicious? And which court defendants are viewed as being particularly high risks? Around the world, states and businesses are using algorithmic systems to ask exactly these types of questions and to make these kinds of predictions. Algorithmic systems are being employed in an ever-growing range of areas and influencing the lives of an ever-growing number of people. Moreover, the quality of automation and algorithmization is steadily improving as information technology becomes increasingly pervasive – more data is collected digitally, and analytical findings are more easily implemented than ever before.

In discussing the use of such systems, it may be easy to rely on conceptual shortcuts, such as asserting that algorithms make important decisions about our lives or that machine-learning systems learn autonomously. However, these kinds of arguments obscure the responsibilities and misperceive the ways in which algorithmic systems are designed and function.

In practice, both the problems and the desired outcomes addressed in an algorithmic process are defined by people. An objective of an algorithmic process might be, for example, to win a game of chess while following the rules of the game. Prior to constructing an algorithmic system, its creators often make decisions regarding the defining of practical and social outcomes of a specific problem to be solved. Take the example of an algorithm designed with the objective of allocating the greatest possible number of students to available secondary school places that takes into account student preferences and demand without adjusting the number of available places or potential shortfalls of placement allocation. Does the proposed algorithm intend to improve a service? Or, does it improve the allocation of scarce resources (secondary school placements) within an existing service (a school system)?

It is currently impossible to establish general socially useful optimization goals that can be applied to the wide variety of algorithmic systems. Indeed, defining and prioritizing societal goals is a dynamic process that must be negotiated each time an algorithmic system is created. Those commissioning, developing or implementing a system that has implications for social inclusion should also facilitate broad public discussions on these issues in a manner that engages those members of society who are potentially affected. The human capacity to work with others to establish joint goals and reach a consensus on desired outcomes has not matched the recent progress in some areas in which artificial intelligence (AI) has been applied.

This working paper addresses the complex interconnections between technology and society. It focuses on summarizing and highlighting potential solutions to understanding, building and regulating new technologies for those working in policy, academia, civil society (e.g., consumer-protection organizations, other non-governmental organizations or activists), businesses and technical developers. It is intended to provide orientation within the field of algorithmic decision-making while providing a toolkit for individuals and groups tasked with making decisions in the design of algorithmic systems. As a result, it contributes to ensuring the human-centered development and use of algorithmic decision-making based on, for example, values of inclusion rather than exclusion. This working paper presents and outlines the spectrum of strategies discussed to date, providing answers to the key questions: What challenges arise when seeking to design algorithmic systems in a socially responsible way (chapters 2 and 3) and what kinds of options are available as we seek to address these risks and challenges (chapters 4 and 5)?

We would like to thank Dr. Ulf Buermeyer, Dr. Andreas Dewes, Prof. Dr.-Ing. Florian Gallwitz, Lorena Jaume-Palasí, Dr. Nicola Jentzsch and Philipp Otto for their critical and inspiring comments.

analytical avenues, as well as any constructive criticism. The Bertelsmann Stiftung seeks to promote the development, formulation and testing of selected strategies. We look forward to hearing from others interested in this subject.

This analysis is part of the Bertelsmann Stiftung “Ethics of Algorithms” project, which examines the societal consequences of algorithmic decision-making. The Ethics of Algorithms project has also published a series of discussion papers, including a collection of international case studies (Lischka and Klingel 2017), an examination of the potential impact of algorithmic decision-making on social inclusion (Vieth and Wagner 2017), an analysis of the impact of algorithmic processes on societal discourse in social media (Lischka and Stöcker 2017), a working paper on detecting problems and solutions in algorithmic decision-making processes (Zweig 2018), and an analysis of how the General Data Protection Regulation impacts the development and use of algorithmic decision-making systems (Schulz and Dreyer 2018).

Ralph Müller-Eiselt
Senior Expert
Taskforce Digitalization
Bertelsmann Stiftung

Carla Hustedt
Project Manager
Ethics of Algorithms
Bertelsmann Stiftung

2 The big picture: Social imperatives for algorithmic decision-making systems

The debate over the use of algorithmic decision-making systems (ADM) is characterized by widespread misunderstanding, a frequent lack of relevant knowledge and a variety of unresolved ethical questions. How broad is software's potential, and what are its limits? How should people handle algorithmically produced predictions? What optimization goals should actors in the public and private sector seek to achieve? And, once these goals are specified, to what extent does it matter whether algorithmic or other decision-making systems are used to realize them?

This chapter seeks to define and clarify a number of aspects relevant to this discussion. Following an introduction to the field's conceptual underpinnings, this chapter offers a structured answer to the following questions: What are the key challenges associated with algorithmic decision-making processes? What conditions must be in place to ensure socially responsible development in use of artificial intelligence and automated decision-making tools, and in this regard, what strategies are likely to meet public expectations?

2.1 Conceptual underpinnings

Algorithms today guide a multiplicity of decision-making processes (e.g., the use of facial-recognition tools by law enforcement agencies) while also enabling new forms of analysis that serve as a basis for human decisions (e.g., resource management in predictive policing). The term algorithm, or algorithmic decision-making, refers to a complex interplay of technical and social organizational functions. The public debate on algorithmic decision-making systems is linked closely to the development of so-called **artificial intelligence (AI)**. This is reasonable given that decisions once reserved for humans are today being automated. While this trend has long been evident in industrial production, the development of artificial intelligence has today brought similar technologies into the administrative, medical and education sectors, for example (Ramge 2018). Any discussion of the opportunities, risks and comprehensive strategies associated with developing and employing socially beneficial algorithmic decision-making tools, however, will require greater conceptual clarity in order to be useful. The following subsection introduces terms and concepts that will figure prominently in the subsequent chapters.

2.1.1 From algorithms to algorithmic decision-making processes

Algorithmic decision-making is, by its nature, based on algorithms. In broader public usage, the term "**algorithm**" refers to a set of precise instructions intended to solve a specific previously defined problem. In mathematics and computer science, the term algorithm is used to refer to predominantly technical instructions expressed in one of many programming languages whereby the *so-called* code specifies a plan by which input data is processed for a specific purpose and then converted into output data. The results of this process provide the basis for an algorithmic decision.

Technologically, algorithms implemented in software constitute:

> *"(...) computer-science tools for the automated solution of mathematical problems. If they receive the necessary information, the so-called input, they reliably calculate the solution to a problem. The mathematical problem defines the characteristics that the corresponding output, and thus the result of the calculation, should have*" (Zweig 2018: 9).

In this context, an algorithm describes one path to a solution that comes into effect after being implemented in computer software and that may, in turn, be used for automated decisions. Such **algorithmic systems (also referred to as algorithmic decision-making systems, or ADM systems)** are employed for specific purposes (see Figure 1). In addition to algorithms, a fully developed software-based ADM encompasses:

- Input and output data,
- An operationalization of the problem to be solved,
- Models for the employment of the algorithms in decision-making.

In short: "*A fully configured algorithm will incorporate the abstract mathematical structure that has been implemented into a system for analysis of tasks in a particular analytic domain*" (Mittelstadt et al. 2016: 7).

For example, if an algorithmic system were to be tasked with determining the relevance of communications within a social network, the diffuse meaning of relevance would first have to be operationalized to allow for a systemic and automated assessment or ranking. The subject of an optimization process must be translated into a machine-readable (mathematical) language and combined with adequate data (i.e., "relevant content"). One way of doing this might be to count the number of positive reactions to a post on social media, with relevant data including the number of comments a given post receives, the number of times it was cited as "favorite," or by other comparable reactions.[1]

Algorithmic systems can be used in a broad range of areas that vary strongly in terms of their complexity. We have long been familiar with automatic ticket dispensers and assembly line robots, for example. But the current debate on ADM systems is primarily focused on the automation of decisions that previously could be made only by people, as in the case of truck drivers, administrators or doctors.[2] Employing algorithmic systems in areas involving knowledge-based work (e.g., the collection, standardization, analysis and processing of large amounts of information) is a complex endeavor, dependent on specific conditions within a variety of non-technical environments. For this reason, when we refer to **ADM processes** in this paper, we are, in fact, referring both to the technical systems and to the ways in which they are embedded in specific social, organizational and cultural contexts.

This process of evaluation clearly entails the multifaceted interactions between people and machines, whereas an algorithmic system's potential for solving the right societal questions with the best (e.g., most effective, least invasive, freedom-of-choice-saving) solution is optimized for the public benefit. Defining a system's optimization goals is not a purely technical process, nor is the selection of data, the task of defining social constructs (e.g., a social-media post's relevance), or the interpretation of the results.

> *"Algorithmic decision-making is always based on certain values and norms. For this reason, the algorithm 'in itself' can never be examined; rather, the way in which it is embedded in a social context must always be considered as well"* (Vieth and Wagner 2017: 11).

The potential of algorithmic decision-making, paired with the power of social-media networks to steer people in a personalized manner, creates novel ethical questions. To take one example, algorithmic systems and data analysis tools could conceivably be developed to identify people experiencing a range of psychological difficulties. Using this information, new forms of social care might be developed and deployed: a system might be able to provide an individual with personalized offers of care and attention, automatically freeze an individual's social media

---

[1] In this case, the underlying model focuses on human interaction and discourse. An alternative model would be to document the amount of time a given user spent on reading a text, or the number of articles a user read completely before sharing a post. A third option for operationalization would be to consider the social or national homogeneity, or the variety of people who share a specific unit of content. All of these conceptualizations would be machine-readable, given the availability of adequate data, but they would differ very much in terms of underlying concept.

[2] For reasons of simplicity and better readability, this publication primarily uses masculine speech forms. However, both genders are meant.

and bank accounts, or even send an automated notification to security services. But which of these alternatives is in fact appropriate? And who will evaluate the effects?

The sociotechnical constructs in which algorithmic systems are embedded also include the procedures for evaluating a system, any options enabling affected parties to register an appeal or objection, and the practical means of correcting errors. This kind of embeddedness, in turn, requires additional questions to be addressed: Are current staff resources sufficient? Are employees trained to deal with an algorithmic system? Can employees properly classify and apply projections? Do the organization's hierarchies and policies allow employees to freely decide whether or not to override the recommendations of an algorithmic system? Examples such as the misuse of the "Strategic Subject List" in Chicago demonstrate that an agency's failures in implementing a system can lead to harmful or biased outcomes regardless of the quality of the algorithmic system used. This case involved a list that helped calculate an individual's risk of being involved in a criminal act. Originally developed for crime-prevention purposes (City of Chicago 2017), in practice, this tool was used by police in their investigation of crimes (Kunichoff and Sier 2017).

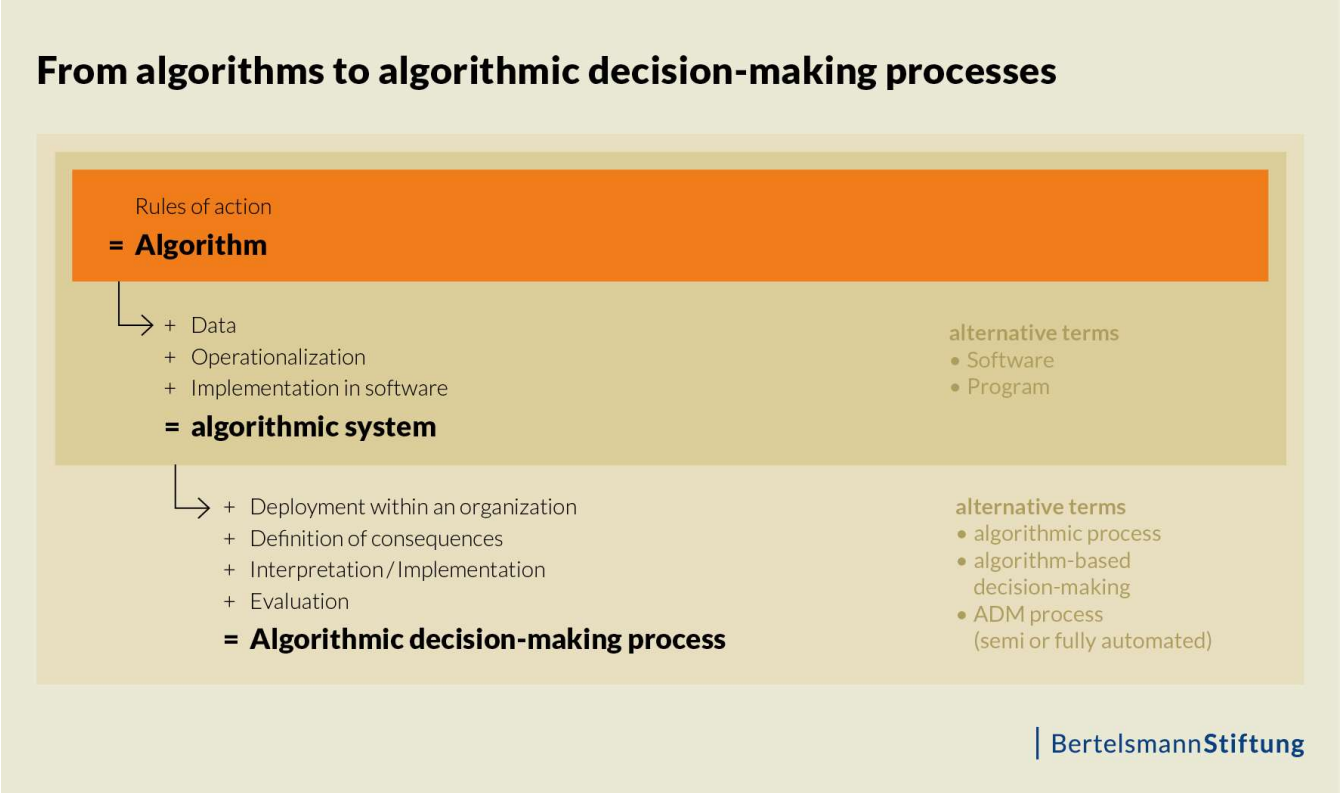
The concept of an **ADM process,** thus, entails a complex interplay between technical and human decision-making. Throughout this paper, our use of the term "**ADM process**" draws on the concept of "**algorithmic decision-making**" as defined by experts in the field (Jaume-Palasí and Spielkamp 2017; Tene and Polonetsky 2017; Wachter, Mittelstadt and Russell 2017). ADM processes are currently used primarily in the business sector. To a lesser extent, they are also used for state activities. Companies use ADM processes to assist with tasks including, for example, applicant selection, the creation of consumer profiles, market segmentation and the personalization of advertising offers. State actors use ADM processes in police work, such as in planning the deployment of police patrols, for automated facial and behavioral recognition (as in a pilot project at Berlin's Südkreuz train station), or for identifying the accents of applicants in asylum proceedings (Bundesamt für Migration und Flüchtlinge 2017). The variety of possible application scenarios is vast.

Today we are confronted by a wide variety of algorithmic decision-making systems, many of which are able to solve a specific problem as well or even better than a human can. For example, an ADM system can successfully win at playing go against a human, identify people in photographs, or transcribe spoken English. In his 1998 paper, robotics researcher Hans Moravec developed his image of a landscape of human competences with this diversity in mind:

> *"Imagine a 'landscape of human competence,' having lowlands with labels like 'arithmetic' and 'rote memorization,' foothills like 'theorem proving' and 'chess playing,' and high mountain peaks labeled 'locomotion,' 'hand−eye coordination' and 'social interaction.' We all live in the solid mountaintops, but it takes great effort to reach the rest of the terrain, and only a few of us work each patch. Advancing computer performance is like water slowly flooding the landscape. A half century ago it began to drown the lowlands, driving out human calculators and record clerks, but leaving most of us dry. Now the flood has reached the foothills, and our outposts there are contemplating retreat. We feel safe on our peaks, but, at the present rate, those too will be submerged within another half century"* (Moravec 1998: 20).

Now, 20 years later, some of these areas are flooded with ADM processes, such as the "go plateau." It remains to be seen which applications will emerge from this, what algorithmic decision-making systems will be converted into ADM processes, what human decisions will be replaced by decisions, and how these decisions will be shaped by future discourse and regulation.

Figure 1



Source: Own illustration

2.1.2 Machine-learning systems and weak artificial intelligence

Looking beyond the complex sociotechnical constructs in which ADM processes are embedded, any analysis of the field must additionally take into account the fact that ADM processes can be based on either machine-learning or non-learning algorithms or algorithmic systems. This distinction is essential as it is of great significance to the question of whether and how ADM processes can be assessed and controlled (see Chapter 4).

For the purposes of this paper, we refer to **machine-learning systems** as algorithmic systems that create problem-solving models with a high degree of automation, a category synonymous with **weak artificial intelligence**. By this, we mean software that learns from data, which "in combination with controllable hardware, [enables] the three-step process of perception, recognition and implementation in an action" (Ramge 2018: 14).

Machine-learning algorithms seek patterns in data, save patterns identified and subsequently apply these patterns to new input.[3] In the course of this process of pattern recognition, an algorithm is trained with specific goals and utilities in mind with the aim to produce effective data analysis within acceptable time and memory-capacity constraints (Russel and Norvig 2012). Humans provide the relevant goal and utility definitions; this constitutes the key difference relative to the so-called **strong artificial intelligence** depicted in science-fiction dystopias. Key

[3] We restrict our consideration of artificial-intelligence technologies here to machine learning tools largely because this technology is currently at the forefront of the field's development (Ramge 2018). Producing even an approximate list of all technologies currently considered to be relevant to artificial-intelligence would fall outside the scope of this working paper.

elements in the implementation of machine-learning processes include **neural networks**[4] as well as various search algorithms that can be progressively trained using a variety of learning styles:

> "*Algorithms search for statistically salient patterns in data. They save this information in various structural forms, for example in a mathematical formula, in decision trees or a neural network. This structure is called a 'model'*" (Zweig 2018:17).

The most important learning styles include:

- **Supervised learning:** A system is trained with data for which the correct classification is known from the beginning of the training process. This specification (the "ground truth") can be based on human (expert) knowledge. An example of this is having a human, whether a depicted person is a man or a woman. In the case of algorithmic systems configured to make predictions, the classification of sufficiently old data sets can take place even entirely without human intervention. An example of this could include, for instance, an algorithm that determines whether a university applicant's profile matches the profiles of other individuals who have successfully graduated (Gallwitz 2018).
- **Unsupervised learning:** In this style, the training data is not classified by a human or old datasets before an ADM process is implemented. The system itself develops its own classification system, or ontology, based on patterns it discovers while an ADM process is running (as per Böttcher, Klemm and Velten 2017: 8). One example of this type of learning process might be a video camera monitoring system in a shopping mall that autonomously recognizes individuals and uses this information to create predictions regarding the frequency of customer visits. If certain characteristics in multiple images of individuals "match" to a sufficient degree, the system assumes that the images must depict the same person.
- **Semi-supervised learning**: This is a mixed form that incorporates elements both of supervised and unsupervised learning. Only a small portion of the data is classified before an ADM process begins, but the type and number of classes are predetermined, as in the case of supervised learning.
- **Reinforcement learning**: The human developer defines success and rewards system actions that achieve this success. Such procedures are today primarily successful where algorithmic systems can act autonomously in closed (often virtual or simulated) worlds. An example of this is when an automated system develops mastery of Atari video-game classics.

The above list of learning styles illuminates one key way in which these mechanisms differ from **rule-based, non-learning algorithmic systems**: Artificially intelligent systems contain feedback loops that measure the effects of their decisions, and integrate these measurements into the results of subsequent decision processes (Ramge 2018). While beyond the scope of this paper, the improvement of results, or the correction mechanisms built into the system, can also be automated to differing degrees. The key takeaway with this point is that a developer does not predefine every individual solution step; rather, the development of the model takes place partially automatically and over time. Nevertheless, the evaluation of the entire ADM process remains of fundamental significance.

Machine-learning algorithmic systems are currently in a phase of rapid development. Many well-known ADM processes are based on **non-learning algorithmic systems**, and thus also fall under the subject matter of this working paper. We define these as systems whose problem-solving models are created step-by-step by human developers. For non-learning ADM systems, the development of an algorithm, the data collection and the creation

[4] "Neural network" is a term used to designate any arbitrary number of interconnected neurons. Such networks form a part of living organisms' nervous systems. Neural networks have also played a critical role in the development of artificial intelligence. Here, they are emulated using data and algorithms; in so doing, digital problem-solving architectures are created that enable information to be processed in a parallel, non-linear and complex way (for a detailed description, see Russel and Norvig 2012).

of a model are all performed in advance. The practical calculation of predefined correlations takes place subsequently as part of the data evaluation process, generating results for further processing. This type of modeling is standard practice within the field of statistics. Examples of non-learning ADM systems include the digital admissions process for state universities (Admission Post Bac, APB) in France (Lischka and Klingel 2017: 25 ff.) as well as the Precobs (Pre-Crime Observation System) software for location-specific burglary predictions (op. cit.: 28: ff).

Distinctions between different learning procedures highlight the essential role played by developers, even beyond the issue of technological implementation. Are developers providing a top-down model, or are they allowing a model to evolve organically within a system? In either case, it is important to keep in mind that developers greatly influence a model's creation. For a top-down approach, this takes place directly through the specification of optimization goals, success criteria and correlations. For a bottom-up approach, the model's development can be indirectly influenced through the selection of the training data.

Finally, it should be emphasized that even in the case of machine-learning algorithmic systems, humans predetermine the goals of the algorithmic decision-making process, the corpus of data to be used as well as the basic models. Considerations regarding the intended areas of use, specific optimization goals and the general conditions in which the system will be deployed, all play a role in the implementation of an ADM process.

In part for this reason, concerns about so-called **strong artificial intelligence** are today unjustified. Strong artificial intelligence would exist if a machine-learning algorithmic system were able to solve problems other than those specified in advance, independently specify its own optimization goals, and autonomously select training data. Strong artificial intelligence exists today only as fiction or as a goal of some research projects (Ramge 2018). Partially automated learning and system complexity – in contrast to total automation – make individual decisions difficult to explain. Problems with rendering decisions intelligible should not be dismissed by pointing to the supposed autonomy of so-called **artificial intelligence**.

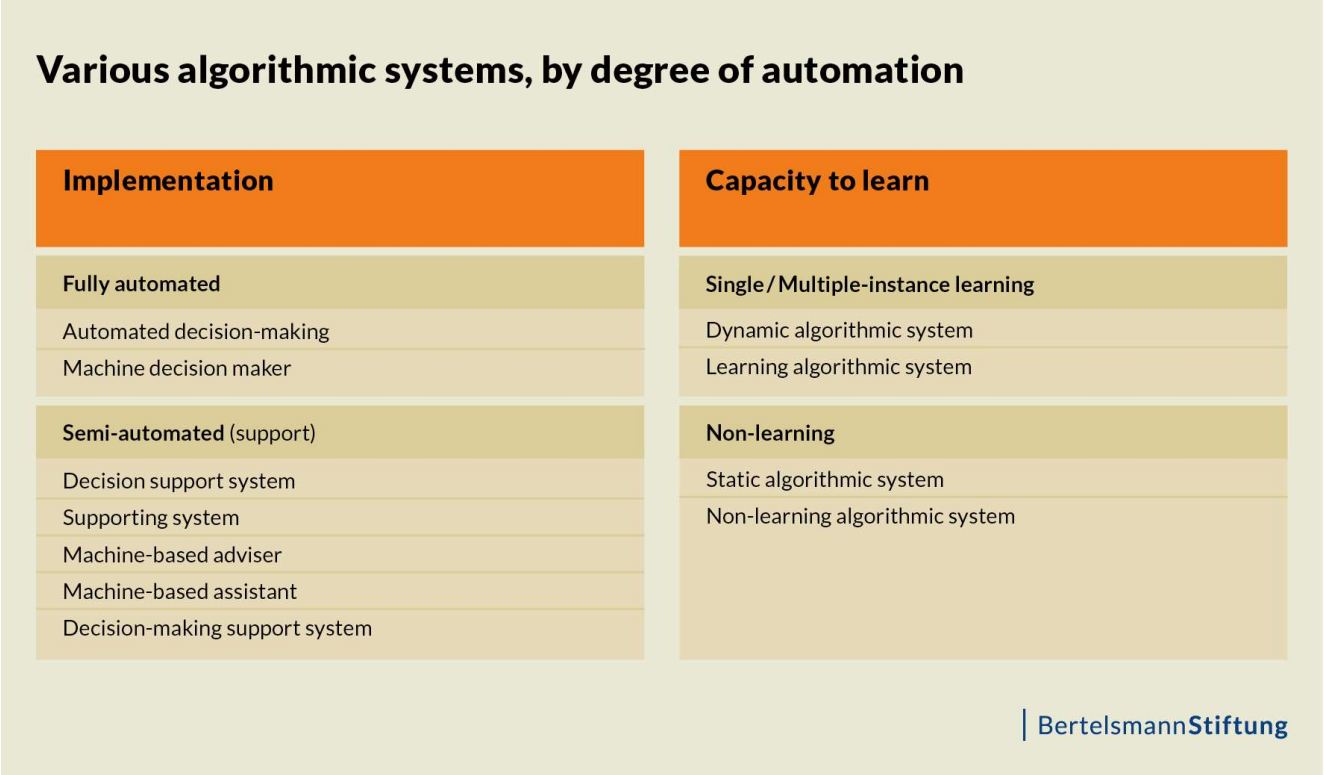
2.1.3 Decision support and automated decision-making systems

In addition to various learning and decision-making processes, algorithmic systems can be distinguished by the degree of their autonomy in transposing outputs into action. A variety of such models exist, including:

- **Decision support systems (DSS systems)**: These advise humans who weigh, make and implement the decisions. This includes software that offers location-specific predictions regarding, for example the probability of burglaries. Using this as a basis, police patrols can make an informed decision as to whether or not they adjust their routes.
- **Automated decision-making systems (AuDM systems)**: This kind of algorithmic decision-making system automatically sets an action into motion on the basis of the output. One real-world example of this approach is a credit-rating system that allows or disallows payment options (invoice, cash on delivery, etc.) based on a person's credit score (Zweig 2018:11).

The implications of this distinction will be further addressed in chapter 3.4

Figure 2:



Source: Own illustration

2.1.4 Social inclusion

Like human decisions, ADM processes can have an effect on social inclusion. For the purpose of this report, social inclusion refers to the equal inclusion of individuals and organizations in political decision-making and policy formation processes. It also refers to the fair participation of all people in social, cultural and economic development. What is at stake here is both participation in democratic processes – and thus political equality – as well as participation in the achievements of a social polity. This includes, “good living and housing conditions, social and health protections, and sufficient and universally accessible education opportunities to integration into the labor market and diverse opportunities for leisure activities and self-realization” (Beirat Integration 2013: 1).

Participation, in this sense, is predicated on all people having access to the minimum level of material resources needed to facilitate their participation in social life. The guarantee of social and political participation, thus, also depends on a “basic equality of fundamental social goods” (Meyer 2016). Elements of this basic provisioning are described in, for example, the Universal Declaration of Human Rights and in the International Covenant on Economic, Social and Cultural Rights (Bundesgesetzblatt 1966). Investments promoting the development of individual capabilities are necessary if we are to facilitate equal participatory opportunities (Bertelsmann Stiftung: 31). It is the responsibility of states and communities to empower each individual to genuinely be able to take advantage of such opportunities.

The stronger the potential impact an algorithmic decision-making system may have in the area of social inclusion, the more rigorously the system must be reviewed. Vieth and Wagner (2017) have outlined one means of comparing algorithmic decision-making systems’ potential impact on participation opportunities. Their framework asks the following key questions as reference points: Are people being evaluated? How much political and economic power does the algorithmic system’s operator have? How dependent are those being evaluated on the outcome of the evaluation? How broad is the system’s reach?

2.2 Social imperatives for algorithmic processes

One need look only to New York City to see how deeply ADM processes have penetrated into contemporary daily life. The city government uses tools of this kind to make decisions about its citizens in a variety of areas. For example, these tools play a role in determining which secondary school students will attend (Tullis 2014), which teachers will receive promotions (O'Neil 2017), where and how often police will schedule patrols and other monitoring activities (Brennan Center for Justice 2017), which buildings have the highest fire-inspection priority (Heaton 2015), and who is suspected of welfare fraud (Singer 2015).

Advocates of such ADM processes invoke a series of potential benefits that can be roughly divided into three areas (see Lischka and Klingel 2017: 37 f.):

- **Consistency:** Algorithmically based projections reliably carry out the predetermined decision logic in each individual case. In contrast to human decisions, software is not swayed by the character on an individual day, and does not arbitrarily apply new, *sometimes inappropriate* criteria in individual cases. Inappropriate criteria, or criteria that are incompatible with social norms, can be excluded from an ADM process from the start, and an ADM system's functions can be documented in detail for each individual case. By contrast, there is ample empirical evidence that humans are inconsistent in making their decisions, and at times show a systematically discriminatory bias (see Kahneman et al. 2016). We can see this, for example, with regard to choosing (or rejecting) job applicants on the basis of foreign-sounding last names (Schneider, Yemane and Weinmann 2014).
- **Complexity management:** Software can analyze and identify patterns within a significantly larger set of data than is possible for humans alone. Certain tasks are not possible or cannot be completed to the same degree in the absence of referencing these patterns or using such tools. Algorithmic decision-making processes can easily tailor their output to individual cases and can adapt to new circumstances more easily than can analogue or manual structures. For example, in the first year after its introduction, the ADM process used for New York's student allocation reduced the number of students not assigned to any secondary school from 31,000 to 3,000 (Tullis 2014). The system additionally took into account the students' preferences, the schools' admission criteria and the number of available places, as the New York Independent Budget Office noted in an independent evaluation (New York City Independent Budget Office 2016).
- **Efficiency:** The algorithmic evaluation of large amounts of data is typically cheaper and faster than the evaluation of comparable amounts of data by humans. A system's decision logic, once developed, can be easily applied to a nearly unlimited number of additional cases. In New York, for example, the Fire Department has praised the efficiency of the centralized algorithmic building-data evaluation process in comparison to an older paper-based procedure that had previously been carried out in 26 separate locations (Heaton 2015). In addition, algorithmic decision-making systems can in many cases deliver faster output than human workers.

Stalder encapsulates the hopes engendered by the prospect of using algorithmic decision-making systems to improve consistency, complexity management and efficiency:

> "*It is precisely an emancipatory politics that, given the real problems, has no desire to withdraw into the illusory world of reactionary simplifications, and which thus needs new methods for seeing the world and acting in it. Algorithms will be among these new methods. Otherwise, we will be unable to manage the steadily increasing complexity of an integrated world based on finite resources*" (Stalder 2017: 1).

The use of algorithmic decision-making systems alone, however, is no guarantee that these opportunities will in fact be realized. New York's use of such tools also provides examples that have not resulted in inclusion gains. The risks to participation opportunities through the use of such systems fall roughly into three categories, distinguished by the following sources of risk:

1. The algorithmic systems' optimization goals
2. The implementation of these goals within specific algorithmic systems
3. The diversity of systems, operators and optimization goals specified within a given area of use

The following chapter describes these three categories, along with the framework conditions relevant to their analysis.

2.2.1 Optimization goals' compatibility with social norms

The New York Independent Budget Office praised the city's algorithmic student-allocation system as it assigned many more students to their preferred schools than did the older procedure. However, the same report expressed doubt as to whether the fulfillment of individual desires was, in fact, the most societally useful optimization goal for such an allocation procedure. Under the new algorithmic student-allocation system, the system systematically assigned students with *below-average* grades to schools with *below-average* ratings. While this certainly fits with the preferences of the students, or of their parents, it disadvantages students from poor neighborhoods, where comparatively low-rated schools and students with below-average grades are disproportionately clustered (New York City Independent Budget Office 2016). Schools can additionally specify preferences, and some schools select the proximity of the applicant's home to the school as the main criterion for selection.

Here, the issue is not the algorithmic decision-making system's efficiency or consistency, but rather the optimization goal itself., Should the algorithmic student-allocation system satisfy individual school preferences in as many cases as possible? Or should it decouple educational opportunities from sociodemographic backgrounds? Both goals are defensible. The fact that an algorithmic decision-making system functions reliably and can be audited effectively by human reviewers says little about its societal utility. The goal pursued by the city – which in turn is reflected in the technology it commissions – should be able to be determined in a consensus-oriented political process that involves as many citizens as possible, especially those who may be affected. In this regard, technical implementation depends on broader societal questions that cannot be answered on the basis of standard criteria.

The design of algorithmic systems that affect participation opportunities virtually always depends on such value-laden goal definitions which are not straightforward and for which there is no consensus: What constitutes a good employee? What distinguishes a relevant journalistic news story? How is an important friendship to be identified? In order to be able to build algorithmic systems designers and developers must first operationalize social concepts of this kind and create them in a way that is measurable and/or quantifiable. Even self-learning systems need optimization goals that are defined by humans. A machine-learning system can carry out an automated search through multiple data sets for factors that correlate with the label "success." However, humans must first decide that successful employees are to be the object of the search and must also decide how to measure success. Therefore, such systems are never autonomous:

> "*It's not enough to improve the quality of a tool, because tools are never neutral. Rather, they reflect the values of their developers and users, or of their clients or research funders. (...) What is viewed as "machine learning" in technical disciplines falls within extremely narrow limits: using trial and error to find the 'best' way from point A to point B, when the criteria for A and B already precisely define what should be seen as the best solution*" (Stalder 2017: 1).

The societal discourse regarding the suitability of certain goals should be codified into generally binding laws only sparingly, as for example, in the case Germany's General Equal Treatment Act for employers. Most instances in which algorithmic systems are used can, with differing consequences, be adapted to a broad range of potential applications. Only on rare occasions are clear, generally shared ideas in place. This means that, for any given algorithmic system, a sufficiently broad societal discussion of the system's goals must be an element of the development process. This is particularly the case if there is no consensus regarding the degree to which these goals are compatible with social norms. This discussion is a crucial means of balancing competing interests in each and every new instance of application, such as balancing the interests of employers, employees and job applicants.

No system should be determined solely by or for one set of interests. Moreover, if this balancing process is to succeed, relevant stakeholders must be involved early on within the development process.

A wide-ranging and inclusive discussion that incorporates the perspectives and experiences of stakeholders is necessary in order to negotiate ADM-optimization goals. Without this discussion, an algorithmic system that learns automatically from training data has the potential to simply perpetuate or reify social conditions of the past in the output it generates. To give a hypothetical example to further illustrate this point: Assume that firm X wants to train an algorithmic applicant-selection system using its current workforce data as training data. The optimization goal is to invite only a certain number of candidates to job interviews, with this selection drawn from those most similar to the most successful 20% of individuals of the current workforce. Because the training data for such a system is based on current employees, an algorithmic selection process has the potential to reproduce the sociodemographic composition of the current workforce's most successful 20% by identifying those from the applicant pool with similar traits or characteristics. Thus, in this example, an algorithmic selection process designed to select the most-qualified applicants has the potential to reify systematic biases in human decision-making. Decisions made by humans are not necessarily fairer than those made by algorithms; indeed, in some application areas, humans are demonstrably more unfair than with their own set of prejudices and, at times, unsuitable criteria. This has been demonstrated in a number of empirical studies. In a German study on employment hiring discrimination, those applicants with foreign-sounding names were less likely to receive an invitation to a job interview than those with typically German sounding names:

> "*In order to obtain an invitation to a job interview, a candidate with a German name must submit an average of five applications, as opposed to seven for a rival with a Turkish name*" (Schneider, Yemane and Weinmann 2014: 4).

A similar effect of reproducing or reifying social norms and biases was presumably at work in the case of New York's algorithmic student-allocation system. Due to the urban area's distribution of wealth and educational levels, the locations with schools rated above average were often in close geographical proximity to the residences of students from comparatively wealthy households. This geographic proximity in turn influenced students' preferences and the allocation results, independently of the algorithmic selection system.

2.2.2 Goal implementation

Good intentions do not guarantee good results. Even algorithmic decision-making systems with optimization goals compatible with social norms can have adverse effects that undermine social inclusion if something goes amiss in its implementation. In order to judge the quality of an algorithmic system, it must be assessed in action.

Implementation quality is, in many ways, associated with the definition of goals. This is vividly illustrated by the video surveillance system tested in a pilot project at Berlin's Südkreuz train station. For this project, an algorithmic system compared surveillance videos with photos of wanted persons with the goal of identifying individuals being sought by the police. For such a program to be effective, at least two optimization goals must be considered: 1) as many wanted persons as possible should be positively identified within the set of all those filmed (high sensitivity); and 2) as few innocent people as possible within the set of all those filmed should be falsely identified as wanted individuals (high specificity). These two goals cannot be maximized simultaneously. Higher sensitivity is correlated with lower specificity, and vice versa.

This point can be illustrated by providing an example: If the recognition system is to recognize virtually all suspects correctly, many innocent citizens would inherently be regularly detained and subject to identification verification. With 160,000 passengers per day at Südkreuz train station and a 1 percent rate of false "wanted" identifications, roughly 1,600 false positives and unwarranted passenger stops would needlessly take place per day.

A system can certainly be optimized in this way (high sensitivity), but whether a society wants this is another question. Thus, as can be seen, implementation is in this case closely bound to the definition and prioritization of

goals and applicable legal restrictions. “The appropriate parameterization of such a system requires a weighing of goods, and may ultimately be a political issue” (Gallwitz 2017:1).

A 2016 study by U.S. based research organization ProPublica provides one of the most well-known analyses of an algorithmic decision-making system’s implementation quality. In their study, the organization examined the quality of algorithmic recidivism projections used by courts in a number of U.S. states. At the point of the assessment, the software had been in use for years, however, the system’s predictive failures had not been previously reviewed or publicized.

The ProPublica study’s key finding was that the nature of predictive failures was different for black and white subjects. The share of black people who were assigned high recidivism projection scores, but who showed no recidivism within two years, was twice as high as the comparable share among white people (Angwin et al. 2016: 2). This research finding set into motion a discussion regarding the system’s fairness criteria and illustrates the necessity of systematic research into the quality of decisions being made by systems with inclusionary effect.

In New York, the lack of intelligibility within an algorithmic teacher-evaluation system prompted a court to bar usage of the software, ruling that the system produced “arbitrary and capricious” results (Harris 2016). Separately, lack of auditability and intelligibility was also among the chief points of criticism regarding an algorithmic system designed to plan New York police patrol routes. According to New York City Councilman James Vacca, the police department was never adequately able to explain to him, as a representative of the people, the criteria and decision logic behind precinct shift planning in the Bronx. “That always annoyed me,” he said. “I felt that I was not being given a lot of the answers I wanted” (Powles 2017: 1).

2.2.3 Diversity of systems and operational models

For individual algorithmic systems, the degree to which optimization goals are compatible with social norms must be assessed, as must the quality of their implementation. However, action at a higher level that comprises the entirety of such systems is also necessary. A diversity of systems and operator models is valuable in and of itself with regard to social inclusion. For the purposes of this paper, diversity encompasses:

- **Diversity of goals and operators**: Different optimization goals in one area of use. This is also associated with a diversity of entities commissioning and operating algorithmic decision-making systems – for example, operators within the public, private and civil society sectors each have different approaches and organizational goals.
- **Diversity of implementation and systems**: Different approaches to operationalization in one area of use.

Ensuring diversity within the universe of algorithmic decision-making can be challenging, largely because algorithmic decision-making systems produce very considerable monopolization pressures. This is particularly the case within the context of artificial intelligence technologies (Ramge 2018: 88). In this regard, two tendencies tend to work together: 1) the network effects inherent in digital platforms and infrastructures, such as hardware and software systems, have enabled companies including Microsoft, Amazon, Google, Facebook, Yandex, Tencent, Baidu and Alibaba to develop oligopolistic data-economy structures; 2) the development of machine-learning algorithmic decision-making systems is based on the existence and use of feedback data, which is primarily available to companies already active in the market. “The more often the [feedback data] is used, the greater the market share they obtain, and the more difficult it will be to catch up” (ibid.).

Those actively seeking to promote diversity within algorithmic decision-making systems should consider the following factors:

- **Scalability related tendencies toward concentration:** Once developed, an algorithmic system’s decision logic can be applied to a wide range and number of cases without substantially increasing the cost of deployment. This enables relatively few algorithmic systems to gain a dominant position in some spheres.

This tendency to concentrate power is greater for ADM systems than for other structures. The lower the diversity of algorithmic systems within a given area of use, the more harmful implementation failures will be for those affected. Similarly, the broader the reach of an algorithmic system, the more difficult it will be for individuals to avoid its procedures and consequences.

- **Reproduction of social plurality and social dynamics:** Social phenomena and concepts, such as the relevance of news stories or the suitability of job applicants can be operationalized in a large variety of ways, depending on context. Moreover, such concepts are subject to societal change. The lower the diversity of algorithmic systems within a given use area, the smaller the space for reproducing society's plurality and social dynamics – in the form of different optimization goals for example.
- **Room for innovation**: When different algorithmic systems are deployed within a field of use, a comparison between them can foster insights regarding impact, sources of failure and alternatives. This constitutes the basis for innovation in implementation, and thus, for social progress.

2.2.4 Creating favorable conditions for socially inclusive systems

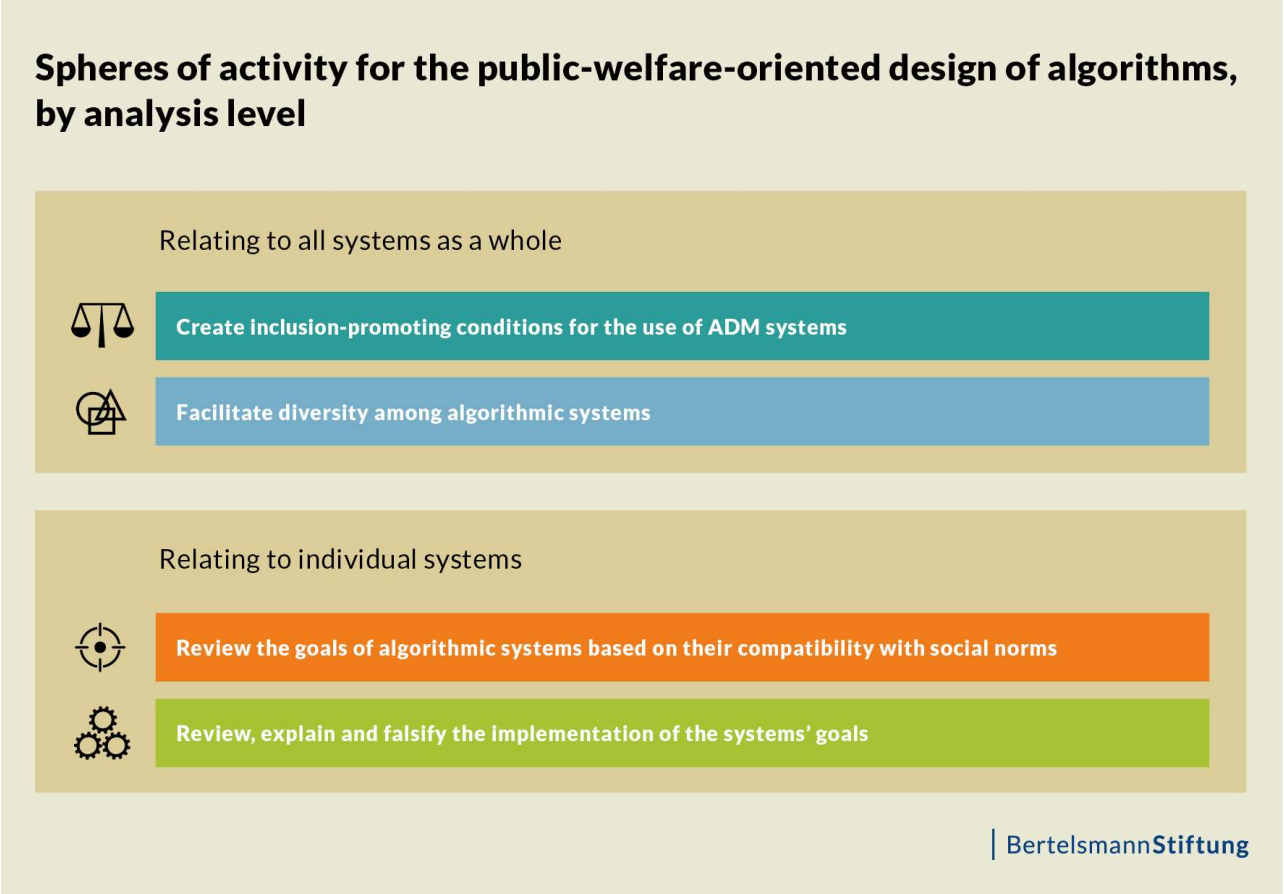
The need for action within the first three fields discussed – optimization goals' compatibility with social norms, the quality of implementation, and the diversity of algorithmic systems and operator models – leads to a fourth field. Competent actors are needed in order to create a framework for positive development. We regard individual and state competence as an essential framework condition in this regard. As cosmologist and Future of Life Institute co-founder Max Tegmark states, societal benefits will not emerge on their own:

> "*I'm optimistic that we can create an inspiring future with AI if we win the race between the growing power of AI and the growing wisdom with which we manage it, but that's going to require planning and work, and won't happen automatically*" (Torres 2017: 1).

Positive, welfare-enhancing design requires state protection and support in the form of regulation. However, the state must also be an active participant in the design process and an enabler. It is the task of the public sector to promote the socially useful deployment of algorithmic systems that promote the general interest – without the field being dominated by special interests like investors and private-sector service providers. Action items here clearly relate to the development of state-level design competencies. One such example is offered by the city of New York, where the city council decided in late 2017 to establish a working group to review the quality of the city's algorithmic systems. In a first assessment of this project, lawyer Julia Powles showed that the working group required competences – both in terms of expertise and capabilities, as well as in terms of legal leverage and authority – in order to fulfill its task. Prowles stated:

> *"There is no readily accessible public information on how much the city spends on algorithmic services, for instance, or how much of New Yorkers' data it shares with outside contractors. Given the Council's own struggle to find answers, the question now is whether the task force will do any better. Can it develop good recommendations, and fulfill its mandate, without the close cooperation of agencies and contractors? (…) The law's second apparent failing is that it doesn't address how the city government, and those who advise it, can exercise some muscle in their dealings with the companies that create automated-decision systems*" (Powles 2017: 1).

Figure 3:



Source: Own illustration

3 Factors to consider: The challenges posed by algorithmic systems

The previous chapter shed light on the conceptual underpinnings of algorithmic decision-making processes and gave an overview of their application areas and the social imperatives related to their use.

This chapter offers a deeper understanding of the complexity in the design, application and evaluation of algorithmic decision-making processes and highlights specific challenges in implementing ADM systems within a societal context. This working paper focuses on processes and structures that influence social inclusion (see Relevance for Social Inclusion Vieth and Wagner 2017). Other ADM processes are therefore only referred to here for the purposes of differentiation or illustration. The greater the relevance to social inclusion, the greater the requirements with regard to safeguarding compatibility with social norms, ensuring that operating principles remain intelligible, and promoting diversity.

3.1 Application area: Are social inclusion issues affected?

This working paper refers to ADM processes affecting the equality of participatory opportunities. This classification itself presents certain challenges:

The concept of social inclusion, which refers to the provision of equal "**capabilities to lead a good life**," focuses on the state's responsibility to ensure all individuals – regardless of their socioeconomic background or affiliation with specific social groups – have equal opportunities to participate in social life. Individuals should be empowered to make use of their individual opportunities on an ongoing basis. Opportunities including equal access to education and employment (Bertelsmann Stiftung 2011) play a key role within this "capabilities approach," and are accompanied by equal access to social protections, health care and recreational activities (German Advisory Council on Integration 2013).

On the one hand, by using datasets that are as large as possible ("Big Data"), algorithmic processes offer the potential to deliver optimally personalized results within each given field of application. This can improve opportunities for participation, such as in the field of education. Here, the use of intelligent software solutions has the potential to support individual learning progresses (Dräger and Müller-Eiselt 2015, Stone et al. 2016). Likewise, the introduction of algorithmic decision-making systems in employment agencies in Poland was intended to improve individualized services for the unemployed (Jedrzej, Sztandar-Sztanderska and Szymielewicz, 2015: 8).

Nonetheless, access to educational or employment opportunities can be impaired by the use of algorithms in some circumstances. Making effective distinctions is one of the essential characteristics of such systems. As illustrated by Barocas and Selbst (2014) through the example of automated candidate selection, the appeal of ADM systems lies in their ability to produce a rational basis for classifying applicants – generally by using a set of criteria that has proved useful in the past. This means, however, that the **intended selection** can easily be accompanied by **unintended discrimination**, as already becomes evident in the definition and operationalization of the goal. For example, if the duration of an employment relationship were used as a key criterion for candidate selection, an algorithmic decision to opt against hiring women would appear logical, because in statistical terms, women are more likely to leave a job after the birth of a child.

In addition to goal-setting and operationalization, the **training and analysis data** utilized by an ADM system also constitutes a many-faceted source of discrimination. In many cases, such data reflects inequalities that are embedded in social institutions. For example, marginalized groups are frequently over- or under-represented (ibid.). When predicting future developments, a dependence on data that is itself discriminatory carries the risk of reproducing the discrimination. Circumvention of this issue through the excision of individual variables that refer to a specific group affiliation – so-called **sensitive attributes** – is not considered feasible, as the attribute will in all probability be reflected in other related data as well. Such related data points are known as **proxy variables** and

might include the postal code of an individual's place of residence, for example. In this case, the connection to the original attribute comes because, like socioeconomic status and ethnic origin, place of residence and socioeconomic status are often correlated.

In addition, the question of discrimination generally applies not only to the representative nature of the data, but also to its **classification and the choice of variables**:

> "*The next biases come from the training data. A data mining system learns by example and must take its training data as 'ground truth,' as that data is the only information the algorithm has about the world outside. A big part of getting the data right is correctly labeling the examples that the algorithm is trained on. The most common source of data for predictive policing algorithms – used in every version of predictive policing in existence – is past crime data, often collected by the police themselves. (...) Reliance on past data is a big problem, though, as accurate crime data often does not exist. There are several reasons for this, but one major one is that the most systematic contact police departments have with 'criminals' is at the moment of arrest. Results after arrest are often not updated. Thus, most research in crime statistics uses arrest data as the best available proxy, even though arrests are racially biased. (…) As a result, a good number of the crime labels may be incorrect, (…) Training data must also be a representative sample of the whole population. The ultimate goal of data mining is pattern-matching and generalization, and without a representative sample, generalizing introduces sampling bias. There are many potential sources of sampling bias. The data can be skewed by past historical practices, for example (…) Another source of discriminatory effect is feature selection*" (Selbst 2016: 17–20).[5]

The challenge lies in ensuring that a dataset is capable of meeting the specific objectives of an algorithmic decision-making system. For example, if the developer of an ADM system wants to rule out any selection of applicants by skin color, it will also be necessary to identify any other attributes in the training data that correlate with skin color (e.g., place of residence).[6] This is only one way of detecting and preventing discrimination through proxy values. Addressing the potential for discrimination in a socially just manner is of crucial importance throughout the entire process of developing and deploying prediction-based analysis systems. Ultimately, this poses a key question: When is discrimination relevant with regard to social inclusion? At what point is it necessary to integrate a certain level of variety or diversity into the **design of algorithmic decision-making systems**?

Vieth and Wagner (2017) have provided an instructive answer to this question. They argue that it is not the **particular area of application,** but rather the **influence and scope of the ADM process** that should be examined in terms of its **relevance to social inclusion**. The crucial variables are a) the political and economic power or market position of an operating entity, b) the existence of alternative decision-making processes and/or product offers, and c) the relationship between algorithmic determination and human decision-making and autonomy of action. In practice, this means that the issue of diversity is of great importance when the ADM process has a major and near unavoidable impact on human activities.

---

[5] The quote illustrates just some of the problems and potential solutions related to the adequate selection and analysis of data. A complete listing falls outside the scope of this working paper. It should be noted, however that false data categorizations can be accounted for in later data processing phases. In addition, any assumptions regarding the data's probability distributions and target values that find their way into the selection of analytical procedures or algorithms must also be taken into account (see. Bayes'sche Verfahren, Russel and Norvig 2012).

[6] Under some circumstances, even variables of this kind – which, taken alone, have no correlation with the target values – can be used to make projections regarding socially sensitive group-related characteristics. In order to genuinely rule out a procedure's ability to extract such information, a training process using the same procedure must be carried out. Here, reviewers will try to predict the sensitive attribute using the original input data. The data can be deemed discrimination-free only if this does not succeed. In addition, the algorithm's outcomes must be monitored over time; the sensitive attribute (e.g., skin color) will again be used in this process.

Nonetheless, the question of what constitutes a **desirable level of diversity** in a society or how to foster participatory justice represents an ongoing **normative challenge** – one that necessitates specific responses in a range of application areas such as health, education and security.

3.2 Goals and evaluation: Who defines and monitors success – and how?

As emphasized previously in this paper, defining goals for algorithmic systems that may affect social inclusion is of crucial importance. These objectives should serve in the interests of the greater social good and include, for example, improvements to health care or education. To this end, the question of whether algorithmic analysis and decision-making systems meet these requirements must be clear and comprehensible. It can be assumed that machine-learning algorithms or artificial intelligence technologies in areas such as health care, public safety or social protection will enable novel insights as well as efficient decision-making and options tailored to specific conditions and needs (Stone et al. 2016). However, these tools entail the risks previously outlined, as well as others (see Chapter 3.5). These additional risks include: system designs that do not conform to the specified goals (e.g., in cases when the essential characteristics of training data fail to correspond to the real data, or when there is a lack of feedback), inappropriate application scenarios, or simply application errors of varying degrees of complexity (Diakopoulos 2016, Future of Privacy Forum, Kroll et al. 2017).

In the academic discourse, the consensus is that artificial intelligence should serve people and society. Nevertheless to date, the field has produced only **general and non-binding guidelines**, such as on the issues of safety, transparency, the preservation of human autonomy, and social justice (Calo 2017; Cave 2017; FAT/ML 2016; Future of Life Institute 2017; Georgieva 2017). The debate over system goals and evaluation is still in its infancy. Guidelines that have been produced have not yet been used for the binding evaluation of disputed cases, resulting in a lack of **concreteness,** as well as a lack of **specific and transparently regulated accountability, assigned to specific groups or individuals**. When considering the field of algorithmic decision-making systems as a whole, one should go about answering at minimum the following questions regarding system goals and evaluation:

- Who defines the goals of automated decision making? Do they correspond to democratic principles? Is it possible to incorporate stakeholders?
- When must the codified goals of automated decision-making processes be made transparent, and to whom?
- Who is responsible for the implementation of automated decision-making procedures? Are the design and implementation in line with the goals of the automated decision-making process, and under what conditions is it possible to deviate from this congruency?
- Who determines the nature and frequency of the evaluation? Should such evaluation be integrated into the system development process?

The answers to these questions are likely to be rather complex. To illustrate this point, we offer an example from the health sector. While personalized diagnosis and treatment of patients through the use development and use of algorithmic systems is possible, balancing the extremely abstract goals of such a system would be difficult to negotiate. Is the objective of such a system to optimize the treatment? Utilize hospital capacities in the most efficient way possible? Reduce costs? Maximize the quality of care in each individual case? (Executive Office of the President et al. 2014; Prainsack 2017; Stone et al. 2016). When designing and implementing algorithmic decision-making systems, different stakeholders routinely make fundamental value judgments of this kind.

3.3 Dynamics and complexity: How has the use of ADM developed?

Many of the most widely known and controversial ADM systems come from debates in the United States, including systems for job-applicant selection, crime prediction and estimating delinquency risk. For these cases, the key criticisms revolve around:

- **Invisibility** (affected parties are generally unaware of the systems' existence),
- **Lack of transparency** (the operations, algorithms and underlying logic are often subject to trade-secret protections),
- **A wide range of influence** (one error can potentially affect a large number of people); and
- **Misappropriation** (systems are used for purposes other than those intended by the developers).

ADM systems have been characterized as "weapons of math destruction" (O'Neil 2016). O'Neill stresses that seemingly neutral ADM systems may be flawed in either their implementation or their application. However, due to the issues of invisibility, lack of transparency and the wide range of influence, such flaws could, in some circumstances, remain undiscovered for long periods of time.

The potential repercussions for affected individuals, particularly those coming from marginalized groups, are considerable. However, if ADM system operators were to provide a transparent system – that is, regarding the underlying data, models and algorithms used – evaluation of such a system would indeed be possible (Zweig 2016). For example, the "risk assessments" produced by the so-called **COMPAS system (Correctional Offender Management Profiling for Alternative Sanctions)** now used in the courts of many U.S. states are based on questionnaires evaluated to produce individual scores. The number of variables underlying this automated decision-making is readily comprehensible. Criticism arises from a lack of transparency regarding the weighting of the variables (the system's mode of operation) as well as flaws within the evaluation process (e.g., with regard to the connection between the delinquency risk and the actual recidivism rate). Particularly high delinquency predictions have been recorded for people of color (Angwin et al. 2016).

The same can be said of a system aimed at identifying and preventing certain known individuals to commit future crimes or locations to experience future crimes, based on existing data of ex-offenders and high-crime areas. These systems are controversial in terms of their effectiveness, correct application and reciprocal social effects, with additional skepticism resulting from lack of transparency (Lischka and Klingel 2017). However, despite this criticism, even these more dynamic processes remain essentially comprehensible both in their functioning and in their effects.

Systems with an unknown and/or unmanageable quantity of data, draw on data analysis technologies from the fields of machine learning or artificial intelligence, or are operationally highly dynamic are quite difficult to evaluate. In these systems, a diversity of data types and data analysis systems (e.g., text, image, activity), can be combined and integrated. **Such complex algorithmic decision-making systems based on highly dynamic interactions** are used, for example, in individual-focused crime predictions based on social-network activity (as opposed to simple individual-focused or location-related predictions, see Selbst 2016). According to media reports (Ferguson 2017), despite a lack of oversight and options for evaluation, these systems are being applied internationally (Dahllof et al. 2017; Dickey 2016, Mateescu et al. 2015). It is difficult to review such a system's compatibility with social norms or evaluate potential impacts for three different reasons: 1) They are based on machine-learning algorithms, 2) they are integrated into highly dynamic systems, and 3) they are controlled by the privately owned platforms that produced them.

Given the scope of currently existing data stores (Christl 2014; 2017) and analytical capacities, the potential of these highly dynamic algorithmic analysis and decision-making systems can only be surmised[7] However one thing is certain: The dynamism and complexity of a decision-making system rises in pace with the complexity of a system's underlying data stores and analysis methods, making system intelligibility and oversight more difficult. As a result, the focus shifts from reviewing an algorithm to a reviewing an overall system – including the suitability of the underlying data, models and input-output relationships (see Chapter 4.2) – along with a system's resulting decisions.

Depending on the scope of the ADM processes, it should be noted that a lack of intelligibility and oversight gives rise to a number of constitutional concerns. The constitutions of liberal democracies embed lawmaking power within the legislature. The power to implement these laws lies with the executive branch of government, which itself is subject to democratic legitimation, that is, free, fair and competitive elections. When an algorithm becomes a "black box" not subject to oversight, this chain of legitimation is broken or, at the very least, eroded.

3.4 Automation: How independent is a decision-making system?

The introduction of automated decisions poses a number of legal challenges. The European General Data Protection Regulation (GDPR), which took effect in May 2018, stipulates that a person generally has the right "not to be subject to a decision based solely on automated processing (…) which produces legal effects concerning him or her or similarly significantly affects him or her." (European Parliament and Council of the European Union 2016). Furthermore, it provides for specific information and transparency obligations for operators of automated decision-making systems, including information on the logic of the decision-making systems, a system's scope and the intended outcomes of a system (see above).[8] At first, this norm appears non-ambiguous. However, questions arise as to whether a decision is automated? And when do these rules apply?

Here, it is possible to distinguish two types of algorithmic systems: decision support systems and automated decision-making systems. Some systems prepare decisions, thereby laying the foundations for decisions to be made by humans. For example, some judges in U.S. states use a software-generated prediction regarding the accused's recidivism risk. Algorithmic systems which serve as the basis for human decisions, such as in the case of the COMPAS recidivism forecasts or in the Precobs software for location-based burglary predictions (Lischka and Klingel 2017: 28 ff.), are here regarded as **assistance systems** or **decision-support systems (DSS)**.

By contrast, other algorithmic systems implement decisions automatically. In the following sections, software is referred to as an **automated decision-making system (AuDM system)** if an algorithmic system issues an evaluation or prognosis, and a software-based mechanism translates this directly into a decision. For example, a piece of software may automatically send out warnings of suspected wrongdoing following a case analysis, as does the Australian Centrelink system (Rohde 2017).

A wide variety of models have already appeared in administrative practice, both with regard to decision-making autonomy and to the associated rights of objection (Citron 2008: 1263 ff.): Examples of highly automated systems without integrated human decision-making include candidate pre-selection on the basis of online personality tests in the Anglo-American sphere, or the allocation of university places in France (Lischka and Klingel 2017). In contrast to this, predictions regarding the delinquency risk of offenders, which are produced automatically by the

[7] Instructive examples can be found particularly in the private sector, for example in revealing statistical correlations between life expectancy and the way a person uses the auto-complete functions of instant-messaging tools.
[8] We reference the General Data Protection Regulation here due to the provision's significant current importance. However, we leave comprehensive legal analysis and interpretation of the issue of decision automation to other studies. We hope the considerations presented here will help inspire discussions of this nature.

COMPAS system in the United States, represent just one of several variables that serve as a basis for an eventual decision by a judge (see above). This assistance system entails a certain degree of human autonomy. The same is true of the software that supports Polish employment agencies in the task of job placement, and which divides jobseekers into different categories corresponding to different support programs (Vieth and Wagner 2017).

While decision-making humans utilizing such decision support systems have the power to deviate from the recommendations issued by the system, studies to date have shown that they rarely do. One major reason is the time required to justify an alternative decision (Jedrzej, Sztandar-Sztanderska and Szymielewicz 2015, Otto 2017). A key question here understanding when and why people make decisions that deviate from the system's recommendations: Does this happen when the ADM system does not share their prejudices? Or if the deviation implies a benefit for the person concerned and no costs for the other party? How do people deal with this shift in responsibility? Does the local work environment, culture and hierarchy permit free decision-making by employees – which may in some cases run counter to the predictions of an algorithmic system – without fear of personal repercussions? Genuine decision-making autonomy must therefore also be considered in relation to the institutional framework. For instance, in cases of doubt, are there incentives for the careful review of decisions, or are penalties for such decisions more likely, for example in the form of negative individual performance evaluations?

These examples show that the question of autonomy is highly important. However, the EU General Data Protection Regulation only covers automated systems where an algorithmic decision leads to a direct action, which encompasses only a small subset of algorithmic decision-making systems. Regulations on information, transparency and rights of objection must also be developed for other system types – either through legislation relating to system operators or through the implementation of procedures that ensure opportunities for human intervention within ADM processes.

3.5 Security: How well is a decision-making system protected against manipulation?

Algorithmic decision-making systems are described as a security risk, particularly when they are used in connection with machine-learning algorithms or artificial intelligence. Some of these security risks have already earlier in this working paper in the context of **appropriate functioning** (see Chapter 2, Amodei et al., n.d.). Although references to a dystopian future are generally considered gratuitous (Ramge 2018), authors such as Stephen Hawking and Elon Musk have described a potentially self-sufficient, artificial intelligence that exceeds human intelligence as a fundamental risk to humanity (Future of Life Institute 2017).

According to Scherer (2016), more practical issues related to the **genesis of risks** are primarily found in specific aspects of the development of algorithmic decision-making systems. While a large part of the development of individual components takes place in an international, as-yet-unregulated setting which includes a vast number of actors[9] the potential of algorithmic decision-making systems only becomes apparent in the interaction between dynamic technologies. This gives rise to risks related to functionality and oversight. Territorial differences in legislation encourage a lack of accountability including potential **manipulations** such as changes to goal orientation, the external shutdown of algorithmic decision-making systems, or external changes to the decision logic (Amodei et al., J. Future of Life Institute 2017, Russel, Dewey and Tegmark 2015).

Unfortunately, policy debates have, to date, paid little attention to these and other issues relating to the security of algorithmic decision-making systems. Rather, debates have tended to focus on issues of standards development

[9] It is currently impossible to specify the level of general risk here, as there has been no scientifically sound survey of the development of individual technologies by individual actors. The production of such material is left to future studies.

and certification. It is often unclear, however, what level of security of algorithmic decision-making systems should be in place for different areas of application. Additionally, few proposals have been made regarding who should set safety standards and how these might be verified (Calo 2017: 14 ff.).

Taking into account the individual components of algorithmic decision-making systems, the following sources of security risks can be identified:

- **False or erroneous data:** This risk increases when there are no regulations on the trading, combination or misuse of data, or on the utilization of data for analysis purposes without the knowledge of those affected (Christl 2014, 2017).
- **Decision-making logic:** For some forms of machine learning, data can be manipulated selectively in order to modify the decision logic. Recent media reports have addressed the development of automated data-manipulation methods for the purpose of influencing the decision-making logic of machine-learning systems (Laskowski 2017). An example from everyday media is the impact and functioning of social "bots," which can influence platforms' trending topics and recommendation systems (decision-making logic) through the dissemination of "fake news" (data).[10]
- **Embedding in the social context:** The development of automated decision-making systems involves risk because it is driven forward internationally in different locations by multiple actors who are subject to different jurisdictions (Scherer 2017). For example, algorithms may be trained at one location using a particular set of data but implemented in software elsewhere. One such example was the test of a facial-recognition software tool that took place at Berlin's Südkreuz railway station. According to the German federal government, it is not known what dataset was used by the system operator to train the algorithms. The generalized nature of the information provided on the analyzing system's automated pattern recognition functions leaves room for a wide range of possible system goals. (Deutscher Bundestag, 19th electoral term 2018: 7). How can it be ensured that such a system will be employed in the interests of the society in which it is used?

What possibilities remain for legislators to prevent uncertainty, abuse or manipulation? How can the development process of algorithmic decision-making systems be fully understood? Who is to be held accountable for unexplainable errors? Who is responsible for systemic risks? And who can investigate and remedy faults in the system? At this time, the topic of security raises more questions than answers. As a result, there is an urgent need for comprehensive analyses in this area. In addition, there are currently legal hurdles in the resolution of security issues (see Chapter 4). These include constraints on adequate security research due to IT-security legislation, copyright law, and commercial and private law.[11]

3.6 Interim conclusion

This section offers a systemic outline of specific challenges that arise in the design, application and evaluation of algorithmic decision-making systems, with relevance to issues of social inclusion.

---

[10] As yet, unaddressed are risks produced by the various analytical procedures themselves. For example, various search algorithms are associated with advantages and disadvantages that depend on the pursued depth, as well as strategies that can and should be carried out in accordance with the time resources and memory capacity available (Russel and Norvig 2012). In the authors' opinion, the field is long overdue for a comprehensive risk analysis that takes into account analytical objectives as well as the various methodologies, and then releases the results in a format suitable for use in societal debates.

[11] The availability of technical infrastructure, computing power and expert knowledge is of considerable importance with regard to the question of who is to investigate and resolve errors. These aspects will not be further discussed here; however, they should be taken into account in the development of state resources (see Chapter 4.4).

The initial focus is on the risk of discrimination, which can be inadvertently intensified and increased by algorithmic decisions. The discussion here shows that more is needed than a sensitive design approach when building algorithmic systems. Far more importantly, we need to underpin this with a normative discussion regarding the degree of diversity that is socially desirable.

However, designing algorithmic decision-making systems that promote the common welfare can be a difficult task due to challenges that arise in the development process. The discussion surrounding such systems' goals and evaluation underscores how different stakeholders are required to make basic value decisions throughout the different stages of system development. This complicates the enforcement of normative guidelines. The reality is characterized by multiple and varied responsibilities, all which must be taken into account.

The pace of technological development, which is leading to ever more complex and dynamic decision-making systems, can make it particularly difficult for humans to comprehend these systems' operations and effects in full. Today, such systems are found predominantly in the private sector, which accumulates sufficient data and has access to sufficient analytical capacities to support their operation. However, the example of predictive policing shows that these complex and dynamic decision-making systems are also gaining a foothold in the public sector. Consequently, solutions for complex systems must be found through cooperative regulatory arrangements.

Beyond this, the question of automation in algorithmic decision-making systems also arises through the use of algorithmic decisions. The prototypical distinction between systems that simply support human decision-making and those that have an immediate impact on individuals may, in practice, become increasingly blurred. This is particularly the case as recommendation systems are optimized for non-interventional processes that, in certain circumstances, undermine the assumption that humans retain the ability to intervene. For this reason, regulations regarding information, transparency and rights of objection must also be developed for partially automated assistance systems.

The development and complexity of algorithmic decision-making systems incorporating artificial-intelligence technologies present particular security risks. These security risks arise partly from the manner of development, which is spread over different areas and stakeholders, and partly from the possibility that individual components might be manipulated. Along with design and implementation errors, these systemic risks underline the importance of developing adequate audit mechanisms for algorithmic decision-making systems.

This chapter has deepened our understanding of the distinctive features and challenges associated with developing and designing algorithmic decision-making systems that promote the common welfare. Drawing on this, the following chapter will outline and systematize possible strategies for the resolution of these issues.

4 The way forward: A panorama of possible strategies

In the following sections, we outline and systematize globally discussed strategies for ensuring that the development of algorithmic decision-making systems is aligned with the public interest. We will focus on the four primary fields of activity previously identified in the chapter on social imperatives related to the use of such systems (see Chapter 2.2). These are:

- Setting goals for algorithmic systems
- Implementation of the goals as the systems are used
- Ensuring a diversity of systems, goals and operators
- Creating favorable general framework conditions for the use of such systems

These four areas of activity require analysis at different levels:

- The issues of goal-setting and implementation must be discussed and reviewed for each individual algorithmic system.
- The diversity of systems, goals and operators can be reviewed and facilitated only at the level of all systems as a whole.
- The framework conditions necessary for the inclusion-promoting use of algorithmic systems implicate each of the above-outlined areas of activity, and thus affect all four. For example, this category may include affected individuals' and system users' skills in dealing with algorithmic processes, as well as the state's regulatory competence.

While the four areas of activity are divided along prototypical lines, individual strategies can certainly apply to more than one area. For example, civil society watchdog organizations represent one group of entities that could review goal implementation in certain systems. However, they also play a significant role in debates over the appropriateness of optimization goals as well as in promoting social discourse on the issue. In each profile, we have noted if such possible overlap exists. In the subsequent classification by field of activity, we have assigned each strategy to the category in which we see the greatest potential for impact. Within each action area, the individual ideas are arranged in the order of their degree of concretization to date, from most developed to the least developed.

The following panorama of strategies shows very clearly that there is no single solution to all challenges associated with algorithmic decision-making. Rather, there is a spectrum of approaches that can contribute to placing algorithmic systems in the service of people and society. Many of the ideas must be further developed if they are to be regarded even as concepts worthy of debate. This working paper is intended to be both a beginning to this process and impetus for further work.

4.1 Ensuring algorithmic systems' goals are compatible with social norms

It is impossible to establish socially useful optimization goals applicable in equal measure to all algorithmic systems. Indeed, defining and prioritizing societal goals is a dynamic process. Each new algorithmic system provides opportunity and impetus to take this process a step further.

> *"From our perspective, addressing the ethical implications of AI poses a dilemma because questions of ethics are about processing and evaluating risks and benefits or acceptable trade-offs in specific circumstances. The area of ethics should not be thought of as prescriptive, but rather as requiring processes for assessing multiple perspectives and outcomes"* (Data & Society 2017: 1).

How can the algorithmic-system development process be designed so as to enable useful goals to be set for systems that may affect inclusion? This key question will be answered as we discuss the strategies outlined below. We focus here on the design of the goal-setting process rather than the establishment of a clear hierarchy of goals that should be pursued in every instance of use.

4.1.1 Documenting relevant interests, stakeholders and optimization goals

Profile: Matrix of interests

Key idea: Various optimization goals, along with their associated interests and stakeholder groups, should be identified, set into relation with one another and documented.

Action area: Reviewing the suitability of optimization goals

Stakeholders: Developers, system operators, all stakeholders potentially affected by the ADM system

Enforcing stakeholders: Developers, system operators, non-governmental organizations, professional associations, state, standardization institutions

Instruments: Process standards, documentation standards

Status: Idea

Anyone developing an algorithmic system must prioritize a system's goals. For example, take a task that appears comparatively simple: A software program intended to allocate patients to beds within a hospital's various stations. Even in the absence of further research, it is clear that this software task could be optimized based on a variety of different goals and their related interests. These goals may include:

- Achieving the best possible hospital occupancy rate
- Obtaining the greatest possible amount of insurance company reimbursement for services provided
- Providing the highest possible quality of care for the patients
- Protecting the reputation of the institution (see Cohen et al. 2014)

Such interests must be described and documented as an algorithmic system is being developed. A kind of **matrix of interests** can be used to depict the relationships between optimization goals, interests and various stakeholder groups. This, in turn, can form both the basis for and part of an impact assessment (see Chapter 4.1.3).[12] By documenting all interests that may either be operative in or affected by the ADM system's development and use, participating actors create a basis for identifying and involving other stakeholders.

Further development of the impact-assessment idea should also address the following questions:

- How can one-sidedness in the description of interests – whether conscious or produced through a lack of knowledge – be avoided?
- With regard to the production of the interest matrix, what actors will be involved in the standardization and quality control processes?
- Could any legal instruments promote the use of such tools to review and balance competing interests (e.g., a documentation obligation that creates standardized procedures for the production and publication of the interest matrix)?

[12] The idea of documentation in the form of an interest matrix was produced in the context of a workshop; the authors thank Udo Seelmeyer and Andreas Dewes for the stimulating discussion.

4.1.2 Informing affected parties regarding use of the ADM system

Profile: Disclosure and transparency obligations
Key idea: Provide affected parties with information about the use and goals of the ADM system
Action areas: Reviewing the suitability of optimization goals, reviewing implementation
Stakeholders: System operators
Enforcing stakeholders: System operators, state
Instruments: Legislation, self-regulation
Status: an initial implementation in the EU-GDPR, idea in early stages of development (application explanations/"counterfactual explanations")

If algorithmic decision-making systems are to be made compatible with social norms, sufficient knowledge regarding their use must be available. Affected parties and those representing their interests can examine software systems' impact, goals and implementation only if they know that algorithmic systems are being used in the first place.

The European General Data Protection Regulation (GDPR) specifies **disclosure obligations for operators of automated decision-making systems** in Chapter 3, articles 13-15, as well as in Article 22 (for partially automated decision-making systems and decision-support systems, see below). This relates to data collected and (see Cohen et al. 2014) processed, insofar as this is used as input for algorithmic processes. However, the regulation also calls for meaningful disclosure regarding: a) the logic involved; b) the scope; and c) intended effects of automated decisions – at least in instances where the decision has legal impact for so-called data subjects or may significantly injure them in any comparable way (for exceptions, see ibid.).

In a study on the human rights dimensions of automated decision-making procedures, the Council of Europe makes a similar call for **effective transparency** regarding specifically:

- The goals of the algorithmic decision-making process
- Any variables used in the software-system's model
- Methodological information (training data, statistical benchmarks, and the amount and type of data underlying the automated decisions; see Council of Europe – Committee of Experts on Internet Intermediaries 2017).

Citron (2008: 1281 ff.) emphasizes that such disclosure requirements are a fundamental part of ensuring due process from both a procedural and substantive perspective.

The demand for **transparency and disclosure requirements** with respect to goals, decision logic and methodology is a key element of algorithm-auditing strategies. Knowledge regarding the scope of the data being used is also essential (Diakopoulos 2016; Zweig 2016).

However, such requirements often remain quite general. For example, the standards contained in the EU General Data Protection Regulation relate exclusively to automated decisions that influence an individual's freedom of action. Moreover, there are also automated decision-making systems that must be assumed to exert considerable influence over our perception of the world. This category includes social networks (Google, Facebook, Twitter), e-commerce companies (Amazon) and platforms for allocating other goods or services (Otto 2017: 18 ff.), for example. Although this potential influence over perception does not trigger any a priori legal consequence, Lenk (2016) argues that the influence over individuals' and organizations' informational spaces nonetheless represents an indirect method of behavioral control and can thus be of societal concern. To give one example, media reports have indicated that Amazon's purchase recommendations for explosives ingredients have in some cases substantially aided suspected terrorists' preparations (Beuth 2017). As a result, experts have also called for a legal

obligation for **transparency regarding the algorithms used in social networks** (Mittelstadt 2016; Pasquale 2010).[13]

Solutions must be additionally found with respect to transparency and disclosure requirements for algorithmic recommendation systems (see Chapter 3.4). Although recommendation systems in application areas such as health or public safety imply opportunities for human intervention, these systems ultimately aspire to an intervention free process, while generally having an effect similar to that of automated decisions. For this reason, the city of New York, for example, makes no distinction between automated decisions and automated recommendations with regard to regulating algorithmic decision-making systems in the public sector. In both cases, decisions are required to be reviewable and auditable (The New York City Council 2018).

However, the dynamism and complexity of new systems remains a challenge with respect to implementing transparency and disclosure regulations. How can statements be made regarding the underlying logic, scope of impact and intended effects if these details change on a regular basis? Google's search algorithm – a complex automated system with more than 200 variables for analyzing, indexing and ranking websites that is modified hundreds of times per year (Pasquale 2016) – offers one example of just how demanding transparency can be.

It can generally be assumed that the interaction of individual software components created by different programmers with only a limited view beyond their own specific piece of work has been a primary cause of unpredictability within algorithmic processes for decades (Passig 2017). Today, algorithmic decision-making processes are based on machine-learning algorithms (see Chapter 2.1.2) whose results are difficult to determine and explain even independently of the specified goals. Moreover, their sphere of application is rapidly expanding. How can society at large understand and help shape the development and results of algorithmic decisions?

To this end, Oxford Internet Institute researchers have offered an instructive proposal they dub "**counterfactual explanations**" – a kind of **usage explanation** of the ways in which algorithmic decision-making systems are being applied. The idea behind this is as follows.

There are a number of reasons that make explaining algorithmic decisions difficult or even impossible. These include technical issues related to the complexity of machine-learning systems, referred to here as "**black boxes**" (see Chapter 3.3). There may, however, also be individual cases for which transparency, while generally desirable, is inappropriate for legal reasons. This includes, for example, the protection of trade secrets, third-party data-protection requirements and the danger that decision systems could be. Yet even in cases for which transparency is difficult, those people affected by algorithmic decisions should have the opportunity to understand and dispute decisions made. Moreover, individuals should also be given guidance as to how they can influence future algorithmic decisions, such as by changing their behavior.

> "*These counterfactual explanations describe the smallest change to the world that would obtain a desirable outcome, or to arrive at a 'close possible world'*" (Wachter, Mittelstadt and Russell 2017: 1).

In the simplest case, such as the granting of loans, an explanation of how the software is being applied might offer additional information on how high an applicant's yearly income would have to be in order for a rejected credit application to be accepted. This method is interesting when many variables or complete variable sets come into play. In such a case, analysis of the variables underlying the algorithmic decision could produce a number of different explanations for the result. According to the concept's creators (ibid.), the most helpful explanation for the

[13] Although regulation of commercial platforms is not a direct focus of this working paper's strategies (addressed largely for the purposes of illustration and delimitation), it should be noted that meeting demands for greater external oversight capabilities or design participation in the context of recommendation systems would require greater transparency than is generally provided today. One possible design option would be the introduction of selection options for users, such as the ability to disable the selection of message contents by relevance in favor of a chronological display.

individual affected by the decision would presumably be one that identifies the earliest possible concrete change. As might be inferred, the outcome of this method is an algorithmically generated explanation of algorithmic decision-making processes.

This approach's novelty lies in the innovative transformation of law into code for the purposes of consumer explanation. However, this mathematical methodology has its problems as well. It proposes an enhancement of algorithmic decision-making systems through the use of algorithmic explanation systems. Yet this conceivably generates the same issues raised in the original problem, particularly in the following areas:

- **Intelligibility**: Can application explanations be audited? Is there source code available that enables oversight, or is this too subject to secrecy constraints?
- **Oversight**: Do application explanations function appropriately and without error? To what extent are they dependent on the availability of an adequate amount of data?
- **Security**: Can the reliability of algorithmic explanations be guaranteed if machine-learning algorithms are being used on a probabilistic basis, rather than deterministically?

In certain areas of use, the idea of application usage explanations constitutes a constructive starting point for the consumer friendly explanation of algorithmic processes. However, while the concept appears to be a promising foundation for future work in the area, it must be further developed in order to earn trust. In addition, potential fields of use should be specified more precisely, with relevant inclusion issues and risks identified for each. One danger here may be that many potential sources of failure may not be able to be identified. In particular, this includes:

- The quality of data in a given database
- Biases in the data structure
- Over- or underrepresentation of certain profiles
- The ADM system's decision architecture (a common source of error)

These problem points all lie within the realm of design and use and may require new oversight systems and procedures.

To date, debates over algorithmic decision-making procedures have largely neglected the issue of how information and transparency requirements can be implemented in a way that is useful for affected parties. One possible starting point is the Chaos Computer Club's idea of the "data letter." This proposal calls for companies, government agencies and institutions to provide regular disclosures to affected individuals regarding the storage of their personal data. The idea even includes data created through processing and combination with other data, such as profiles, assumptions about preferences, and internal customer-class categorizations (frankro 2010). The possibility of a class-action lawsuit right is addressed in Chapter 4.2.4.

4.1.3 Reflecting on and documenting expected outcomes and effects

Profile: Algorithm impact assessments

Key idea: Provide a public overview of the use and goals of ADM systems, which can then be used as the basis for further measures (e.g., an algorithm-focused technical oversight board modeled after the German Technischer Überwachungsverein (TÜV; Technical Inspection Association); stakeholder-participation processes; general system reviews)

Action areas: Reviewing the suitability of optimization goals, reviewing implementation

Stakeholders: Policymakers, system operators

Enforcing stakeholders: System operators, technical associations, state

Instruments: Legal requirements, self-regulation, process standards, documentation standards

Status: Idea, early prototypes (FAT/ML), proven examples from other areas (e.g., construction industry)

Transparency and disclosure obligations are intended to give affected parties an overview of the bases of the decisions affecting them. This, in turns, provides an opportunity to contest and perhaps change the decision. But how does this overview function at the societal level?

Impact assessments for algorithmic decision-making systems could be one important means of offering the public an overview of the systems being used. According to computer scientist Ben Shneiderman (2016: 13539), the use of flawed, discriminatory or harmful systems can be stopped only with such an overview. Schneiderman proposes that system operators be required to produce an impact assessment that could then form the basis for an independent oversight review.

These proposals are modeled after the environmental impact reviews and statements widely used in the European Union and the United States, which provide information regarding affected stakeholder groups and potential consequences in circumstances such as construction projects, for example. An "**algorithm impact statement**" could provide the public with an overview of a number of key issues, including:

- The goals of an algorithmic decision-making system
- The quality of the data inputs
- The expected results (ibid.)

Deviations from an intended mode of functioning could be identified in this way, for example. In discussions in Germany, a similar idea has been discussed under the rubric of a "Beipackzettel" (essentially, the instruction leaflet included inside a package). As might be inferred, this would be a document identifying the "area of use, model-related assumptions and societal side effects" of an algorithmic system" (Zweig 2016). Any further development of the idea of impact assessments should also address these issues:

- Expected individual-level side effects
- Expected quality (e.g., rate of failure)
- Reasonable quality-control measures during operation (e.g., a comparison between training and real data in the case of automated machine-learning systems)

The idea of a so-called **discrimination impact assessment**, proposed by legal scholar Andrew Selbst (2016) in the context of prediction-based police work, commonly referred to as "predictive policing," draws inspiration from similar sources. For predictive policing, however, the review would focus both on an algorithmic system's effectiveness and its potential discriminatory effects. The proposal involves comparing different algorithms and models within the same application area. This would enable public involvement in the selection, design and development of algorithmic decision-making systems. The opportunities and risks associated with prediction-based technologies could be the subject of longer-term debate. Moreover, trust in the work of the police and security services would be enhanced.

The demand for some kind of **social-impact assessment** for algorithmic decision-making systems has also been expressed by the **Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)** expert network. In FAT/ML's "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," the authors take both the decision-making system and its development and application contexts into consideration. While discussing a social-impact assessment of this nature, they additionally call for data operators to disclose information regarding internal lines of responsibility; documentation, transparency and objection obligations; and algorithm-auditing measures (FAT/ML 2016).

For certain areas of use, the designers of algorithmic systems could also be legally obligated to produce such an evaluation in the form of a **risk forecast**. This latter proposal is offered by Martini:

> *"Anyone who implements algorithms in their software that have the potential to produce significant personal – and particularly discriminatory – risk, should be required to create a risk forecast. In doing so,*

> *they must analyze and disclose the degree to which the digital system endangers constitutionally protected values, and what technical, organizational and legal protective measures are envisioned in order to avoid breaches of the law"* (Martini 2017: 1022).

Under such a model, system operators would be urged to provide the relevant information. Various models ranging between self- and co-regulation are conceivable; for example, system operators could be subjected to legal obligations, or could make a voluntary commitment to abide by certain standards.

4.1.4 Ensuring broad stakeholder participation in the development and deployment phases

Profile: Expert committees for stakeholder participation

Key idea: Institutionalize stakeholder-participation processes in the ADM development, implementation and evaluation phases

Action area: Reviewing the suitability of optimization goals

Stakeholders: Developers, system operators, users, all those potentially affected by the ADM system

Enforcing stakeholders: Developers, system operators, technical associations, standardization institutions

Instruments: Process standards, participation mechanisms

Status: Idea, possible models in other areas (expert committees/"IT review boards")

The development and use of algorithmic decision-making systems include value judgments that are built into every system's design from the start.

As the previously discussed example of personalized diagnosis and treatment mechanisms (see Chapter 3.2) shows, this can entail a clash between **contradictory interests**. For this reason, legal scholar Danielle Keats Citron (2008: 1288 ff.) calls for affected stakeholders to be involved in the process of developing inclusion-relevant algorithmic decision-making systems. This is necessary because the development of such systems includes a **transposition of political and societal goals and agendas into code**, she argues. In this sense, it becomes a kind of legislative undertaking. Consequently, the process must be accompanied by the publication of any rules being set, and also by opportunities for public debate. Otherwise, the democratic process would be undermined.

This demand seems to be of fundamental importance **in areas of application that are relevant to inclusion**. In the United States, there have been positive experiences in this regard in areas such as social security. One concrete option for institutionalizing participation lies in the **formation of expert committees** ("information technology review boards") that would allow experts and the public to participate in system development, implementation, use and evaluation (ibid: 1312).

Promising **stakeholder-participation models** can also be found in the bioethics and medical ethics spheres. In the context of genetics research, alternative solutions to ethical problems have been crafted at various societal levels. According to Cohen et al. (2014), one possible approach would be the **integration of patient representatives** into organizations entrusted with the development and use of algorithmic decision-making tools in the medical sector. There have been positive experiences with allowing the use and exploitation of genetic-material databases overseen by a **trustee** (typically an individual or a group of affected parties). However, the trustee's rights in instances when the algorithmic decision-making system is misused or otherwise leads to problems would have to be clarified. An alternative to a model of this kind, the authors note, would be to establish **community-level expert committees** in which stakeholders such as patients, doctors, hospitals and data scientists would jointly develop recommendations (ibid.).

This overview indicates that there are **numerous possibilities for institutionalizing** stakeholder participation. As previously noted in the context of impact assessment (see Chapter 4.1.3), any such measure must consider

the development of algorithmic decision-making systems as well as their concrete implementation, use and evaluation.

Any further development of the idea must also address the issue of how the potentially affected public – conceivably a very large population – can be effectively integrated into these processes. What stakeholders could effectively translate the technical issues at hand and thereby render them comprehensible for a broad public? How can design details that require technical knowledge to discuss be translated into issues of societal relevance?

Regulations of this kind are instruments of **self- and co-regulation**.

4.1.5 Establish industry-wide ethical standards

Profile: Development of professional-ethics codes and institutions
Key idea: Codify process-related quality standards for the design of algorithmic systems in order to ensure minimum levels of diligence, explainability and impact assessment, for example.
Action areas: Setting goals, implementation
Stakeholders: Developers, companies, researchers, universities, professional associations, non-governmental organizations
Enforcing stakeholders: Professional organizations
Instrument: Self-regulation
Status: Ideas, first steps, possible models in other areas (medicine, medical research, journalism, social work, psychology)

A number of professions have produced principles for judgments relating to people's well-being, while additionally establishing institutions that assess specific cases on the basis of these principles. The Hippocratic oath is perhaps the most well-known basis for such a professional ethic. Updated in the form of the World Medical Association's Geneva Declaration, this is still placed at the head of the German doctors' professional code of conduct today (Bundesärztekammer 2015: 2). When discussing the evaluation of algorithmic systems' compatibility with social norms, many experts recommend developing an equivalent to the professional ethics standards and institutions that exist in areas such as medical research. The AI Now Institute offers one such example of recommendations from the civil society sphere:

> "*Ethical codes meant to steer the AI field should be accompanied by strong oversight and accountability mechanisms. More work is needed on how to substantively connect high level ethical principles and guidelines for best practices to everyday development processes, promotion and product release cycles*" (Campolo et al. 2017: 2).

As this idea is further developed and implemented, conceptual questions in three broad areas must still be answered:

1. Who is being addressed by the code of professional ethics?
2. What should be included in the code?
3. How will the principles gain recognition, be translated into actual practice and updated?

1. Who is being addressed by the code of professional ethics? A variety of professional groups are involved in developing algorithms, collecting data and developing models, as well as in implementing, using and evaluating individual systems. Data scientists are certainly involved in many of these procedural steps (Zweig 2018), but not necessarily in all of them. And what about product managers or statisticians, for example? If a code of professional ethics is to address all those involved in the process of developing and deploying such systems within their ultimate social context, ensuring that the principles are recognized, regularly updated and viewed as binding will

be a challenge, particularly given the lack of a common foundation such as a professional association. Furthermore, abstract concepts aimed at everyone involved in developing and implementing algorithmic systems are likely to be more difficult to establish than would be principles within a single profession with existing professional associations and established training paths. The weaker a group's self-perception as a profession in any given sphere of activity, the more difficult it will be to create an ethic that participants view as relevant to their professional self-conception. How can such a document nonetheless be made binding? These are among the implementation issues that must be addressed in the course of developing this idea further. The proposals made thus far offer no suggestions in this regard.

2. What should be included in the code? Algorithmic systems can be used in a huge variety of different sectors, ranging from medical diagnostics to the justice system or personnel selection. This breadth of use constitutes a particular challenge in developing a professional ethic. For example, it means that optimization goals and the consequences of implementing algorithmic processes' system outputs are not determined by algorithm developers alone, and indeed may not even be reviewed by them. This is different than in the case of professions such as medicine, journalism or social work.

The question of whether optimization goals are compatible with social norms extends beyond the basic issues of system design and the selection of data. What consequences does the system's use have for the individuals being evaluated? What foreseeable consequences does its use have for collective goods? What alternatives exist? For societally relevant systems, these questions cannot be answered solely by data scientists, product managers, implementers and the others involved in system development. A variety of reference points and instruments are needed here in order to enable societal dialog on the issue and set it effectively into motion. One strategy in developing a code of professional ethics is to facilitate this dialogue. In cases where algorithmic systems affect social inclusion, their goals, design and functioning must be subject to societal oversight and political processes. This feedback can also drive the development of the code of ethics:

> "*It was ultimately the civil society discourse that drove the bioethics search for norms in the medical field, and particularly their foundation in medical education, practice, research and institutions, for example in the design of clinical ethics committees*" (Frick SJ 2018: 103).

The participants in the development of the system have an ethical responsibility to foster social processes for coming to a consensus on the issue. In the following, we will refer to this approach as a **process-based professional ethic**.

The Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) initiative's Principles for Accountable Algorithms move in this direction. The authors state that this goal of a process-based professional ethic would:

> *"(…) to help developers and product managers design and implement algorithmic systems in publicly accountable ways. Accountability in this context includes an obligation to report, explain or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms"* (FAT/ML 2016).

The authors additionally posit five principles for shaping algorithmic systems using these maxims:

- **Responsibility**: Create opportunities for complaints and appeals and make certain these are publicly visible.
- **Explainability:** Ensure that decisions can be explained to end users and other stakeholders using non-technical terminology.
- **Accuracy:** Identify, document and record sources of failures and uncertainty in the system so that outcomes can be understood, and mitigation measures developed.

- **Auditability:** Enable interested third parties to test, understand and review the system, for example through suitable application programming interfaces (APIs), detailed documentation and terms of use that allow such examination.
- **Fairness:** Ensure that algorithmic decisions do not lead to systematically and unfairly different output for different demographic categories (sex, national background, etc.).

These ethical principles are very abstract and are not exhaustive. For example, aspects such as data protection and guidelines for human-subject experimentation are lacking, in part because the U.S. legal culture from which this proposal arises is different than that in Germany or the European Union. Despite these outstanding gaps, this process-based professional ethic approach (FAT/ML 2016) is quite promising.

The Association for Computing Machinery (ACM), a computer-science association, has proposed a comparable set of process-based professional-ethics principles (ACM 2017).

Unlike the process-based approach, an **outcome-based approach to professional ethics** seeks to incorporate a system's achieved goals – that is, the "ultimate impact of an AI system" (Campolo et al. 2017: 33) – in the evaluation. An example of this would be the three general principles of Ethically Aligned Design, a concept proposed by the international Institute of Electrical and Electronics Engineers (IEEE), a professional association: "1. Embody the highest ideals of human rights. 2. Prioritize the maximum benefit to humanity and the natural environment. 3. Mitigate risks and negative impacts as AI/AS evolve as sociotechnical systems" (ibid: 15).

Any further development of the idea of professional-ethics codes should also address these questions: How can the requirements that make up the content of the code, as well as the means of ensuring they are binding, distinguish between private and state technology use? The state is directly mandated to ensure equality and inclusion, while legal requirements for the private sector are less stringent. Should codes of ethics take this differentiation into account? Should the codes consider systems' potential impact within the field of use? This latter issue appears reasonable, at least in areas where private operators are fulfilling public functions (e.g., in the creation of infrastructure for social discourse).

3. How can the principles of a professional code of ethics be made binding, and how will they be updated? Experts often propose integrating professional ethics into training programs in the mathematics, data science, machine learning, computer science and other relevant fields. The challenges here begin with specifying which specific fields should treat this content as mandatory. One difficulty here is that many professional groups are involved in the development of algorithmic systems, while data scientists, for example, have no formal training program. Given this constraint, the model used by medical schools, in which courses on medical ethics are mandatory, cannot be readily adopted. For its part, the Royal Society (2017: 12) suggests "relevant training in ethics" for "postgraduate students in machine learning." The Obama government's expert commission on artificial intelligence, by contrast, recommended a thematically and organizationally broader approach:

> "*Schools and universities should include ethics, and related topics in security, privacy and safety, as an integral part of curricula on AI, machine learning, computer science and data science*" (Executive Office of the President et al. 2016: 34).

With regard to training, the next step is to specify who should learn what. Some mechanism is also needed to help people no longer in school gain familiarity with this code of professional ethics. This is particularly relevant given that many people working as data scientists come from the natural sciences. Another area that needs fleshing out is the question of how a code of professional ethics, when applied to new cases, can expand or update its principles accordingly. Here too, a number of fields offer possible models. For example, case review in interdisciplinary teams, as performed in the context of clinical ethics committees, is a promising practice. Similarly, the German Press Council has an institutionalized committee-based case-review mechanism, as well as a clearly regulated process for examining potential updates to the German Press Code. Such examples of successful

codes of professional ethics in other fields must be analyzed, and where possible transposed to the area of algorithmic systems.

4.2 Reviewing the implementation of goals within systems

The preceding chapter examined strategies for ensuring that the goals set for algorithmic decision-making systems are compatible with societal norms. Of course, there is no guarantee that these intended outcomes will be achieved, or that unintended outcomes can be excluded. Especially in the case of machine-learning systems, the issue is not only that programming failures or bugs may occur, but that unexpected interactions between the elements of such complex systems may arise. In addition, it is possible that individuals could simply fall victim to the statistical workings of probabilistic decision-making systems. This is particularly true for predictions that consider not only the individual's behavior, but also that of friends, family members or network contacts, as is currently common practice in predicting recidivism. For this reason, algorithmic decision-making systems must be auditable, and must be subject to appropriate oversight (FAT/ML 2016; Future of Privacy Forum 2017; Web Foundation 2017). Indeed, this kind of activity is essential to assessing and evaluating systems currently in use.

Thus, this chapter addresses options for auditing algorithmic decision-making systems. It will initially provide an overview of technical and institutional alternatives for analyzing algorithms (auditing), followed by a presentation of strategies for ensuring that algorithmic decision-making systems are based on adequate data. The question of who carries out the review, furnished with what capacities, is crucial. In addition to an internal audit conducted by the system operator itself, other possible options include the establishment of civil society watchdog organizations, or the creation of a publicly supported institution tasked with approving and overseeing algorithmic decision-making systems. Legal restrictions that arise in the course of reviewing algorithmic decision-making systems, along with corresponding solutions, complete our survey of the comparatively narrow field of oversight options.

4.2.1 Developing methods for reviewing system implementation

Profile: Algorithm auditing
Key idea: Develop technical and procedural methods for reviewing the functioning of algorithmic decision-making systems
Action area: Reviewing implementation
Stakeholders: System operators, system developers, researchers, non-governmental organizations, regulators
Enforcing stakeholders: System operators, researchers, state, regulators
Instruments: Technical procedures, data-access standards, legislation, process standards
Status: Ideas, early prototypes, models in other areas (qualified transparency, for example in the context of financial audits)

A growing number of proposals have raised various possibilities for auditing algorithms. This involves examining both the functionality and the impact of algorithmic decision-making systems (Mittelstadt 2016).

Algorithm auditing or – as it was previously termed – algorithm analysis is a fundamental mathematical discipline. For simple algorithmic decision-making systems, a number of development and deployment phases serve as possible points of failure (see Zweig 2018: 17), including:

- Algorithm design and implementation
- Methodology choice and operationalization (data collection and data selection)
- Construction of the decision system
- System training, for machine-learning systems
- Deployment within an embedded societal context
- Revision of the decision-making system

The authors see promising possibilities in the review of entire systems (**ADM system reverse engineering**), which depends upon system operators providing transparency regarding all system components (Zweig 2016).

Reviewing algorithmic systems can also be more complex, however. Indeed, for as long as 50 years, not just algorithms but complex software systems in general have threatened to outstrip human understanding and oversight capabilities. This is due particularly to the following factors:

- The number of programmers that work on individual aspects of a given algorithmic system
- The unpredictability of the interactions between various system parts (Passig 2017).

System complexity rises with the number of mutually interacting system elements, a situation frequently encountered with the integration of artificial-intelligence technologies. Thus, the above-described procedures represent only one possible strategy for cases in which the algorithmic system is comparatively simply constructed. This category includes the U.S. court system's delinquency-projection tool (Lischka and Klingel 2017), for example, or the classification of unemployed individuals for the purposes of developing adequate services, as is currently done in Poland (Jedrzej, Sztandar-Sztanderska and Szymielewicz 2015). In addition, the individual components – the problem to be solved, the corpus of data being used, and the underlying models and algorithm – must be disclosed, at least to the actors responsible for the review.

Complex algorithmic decision-making systems, whose system elements interact dynamically and include artificial-intelligence technologies, are currently used in areas such as the creation of consumer profiles, market segmentation, and the personalization of information in recommendation systems. These applications fall outside the scope of this survey of inclusion-relevant decision systems. By contrast, the use of artificial-intelligence systems for border controls, as have been deployed in Australia, certainly falls within the scope of this study. Nonetheless, methodology remains of crucial importance, with an audit of these systems demanding a variety of methods. Starting points here include approaches inspired by traditional field testing or **input/output analyses** (Sandvig et al. 2014). These focus on the systematic analysis of the data underlying an algorithmic decision-making system, as well as on the system's results.[14] The analytical focus thus shifts from access to the algorithms to access to system inputs and outputs. However, because these are in most cases proprietary systems, direct access to these data streams is also generally restricted. External analysis is subject to legal restrictions (see Chapter 4.2.4).[15]

Automated procedures such as **web scraping** and data accumulation by bots generally suffice for the data-collection portion of such analyses. Nevertheless, because these methods are quickly reaching their legal limits, various civil society institutions are experimenting with **data donations**, a kind of **collaborative research** into the functioning of large platforms (Kitchin 2016; Sandvig et al. 2014). The AlgorithmWatch initiative, active during the 2017 German federal elections, offers an example of collaborative research into the Google search algorithm. This project was focused on the question of how strongly search results are personalized, or how strongly they are algorithmically adjusted for each searching person. In pursuing this goal, users of the Spiegel Online media platform, daily newspapers and social-media accounts were asked to donate their search results for 16 search terms, primarily relating to leading politicians and parties. This was technically accomplished through the use of a so-called **plug-in** – a browser extension that automatically transferred the relevant data. With up to 600 data donors per day, there was only limited indication of strong personalization. However, a certain regionalization was

[14] In this sphere, there are numerous testing systems able to review software on the basis of its correctness. According to experts, no single procedure is sufficient on its own for the review of algorithms. However, tests with real data, extreme-value tests, fuzzing and static analysis all represent starting points (Dewes 2018).
[15] So-called property-based algorithm testing constitutes a special form of input-output analysis. This involves analyzing or characterizing an algorithm on the basis of artificially generated data. This can be interesting particularly for system operators examining new application fields (Dewes 2018).

evident – that is, an adjustment of search results to reflect local characteristics such as regionally active groups (Krafft et al. 2017).

The study's authors have acknowledged that it was not representative. Nevertheless, it was an innovative project that offers considerable room for further development, and which provided a first approximate look at the functioning of a non-transparent algorithmic decision-making system that influences public perception. However, the study also demonstrated that alternative data-collection methods pose significant problems with regard to representativeness and speed. This in turn argues for the creation of a requirement mandating ADM-process operators to allow standardized research within certain fields of application.

The development of **algorithms that generate explainable variables and models** (Diakopoulos 2016; Gunning 2016) represents a wholly different approach to the review of complex decision-making systems. The application explanations described above (see Chapter 4.1.2) fall into this category of method. Based on mathematical models, these tools provide information on how complex algorithmic decisions could be changed in individual cases. These approaches represent innovative and constructive strategies, with considerable ongoing development evident.

However, the quality of algorithmic decisions cannot be guaranteed simply through a review; for this, it is necessary to set **technical standards** as well. One such example is the visa-award system in the United States, which allocates access opportunities for applicants using a raffle system. The procedure includes a diversity goal; that is, a certain number of visas are meant to be awarded to people from origin countries that would otherwise be under-represented. But who guarantees the system's randomness? The award system is difficult to review, as both the data, on data-privacy grounds, and the algorithm itself, for reasons of potential manipulation, are kept secret. Given this fact, **procedural requirements** are intended to secure the quality of the decision process; these include factors such as **encryption technology** and certain forms of **information-science tests** (Kroll et al. 2017).

This overview of basic approaches to the issue of algorithm auditing shows that this is a broad field that is not easily regulated through legal obligations, as technical complexity and legal hurdles sometimes stand fundamentally in the way of effective oversight. Nevertheless, some observers continue to demand that algorithms generating automated decisions be subject to regular and institutionalized oversight and certification (Citron und Pasquale 2014; Council of Europe – Committee of experts on internet intermediaries 2017: 32; Diakopoulos 2016: 26). A number of potential solutions appear particularly promising, including:

- Research projects that collect relevant data in roundabout ways in order to conduct input-output analyses, and thus also help – at least to some extent – to equalize the informational balance-of-power relative to the large system operators.
- **Technical measures** that enhance the explainabilty of algorithmic decision-making systems, and thus help ensure fairness.

While the analysis of ADM systems themselves is vital, examining the way they are embedded in ADM processes is also of consequence. This includes the forms of input used, interfaces with other systems, and so on.

Of course, it should also be noted that algorithm analyses do not necessarily need to be made public against the system operators' wishes. Such reviews can also be conducted by closed expert groups that commit to confidentiality or secrecy standards regarding the systems or system elements (**qualified transparency**), thus going beyond formal reporting requirements. This represents a promising possible solution particularly in cases where protections against manipulation or individual data-privacy factors are of great significance.

In order to better understand the opportunities, constraints and necessities associated with various algorithmic decision-making systems, it would be useful to create a **classification of the different kinds of systems**. This could take into consideration:

- The type and complexity of the automated decision-making system
- The operator (e.g., public or private sector) and area of use
- Potential risks to affected individuals (Tutt 2016)

This classification can be used in allocating resources for the audit of algorithmic decision-making systems that demand greater attention because of their scope or potential risks, or where it is especially important to prevent unexplained failures (ibid.). In such cases, it may be possible to reach agreement on **institutional solutions** that allow for an audit and simultaneously protect reasonable interests (**qualified transparency**) – for example, through an external expert group, an agency or independent institution.

4.2.2 Improving and documenting data quality

Profile: Data-correction standards and data-quality institutions
Key idea: Develop better methods of ensuring that data is current, accurate and complete
Action area: Reviewing implementation
Stakeholders: Affected parties, system operators, researchers
Enforcing stakeholders: Policymakers, system operators, state
Instruments: Legislation (implementation), development of standards
Status: Ideas

A hypothetical example may help illustrate the importance of the data used within algorithmic systems. Say that company X wants to determine whether there is a correlation between its employees' social backgrounds, their CVs and their career arcs within the firm. To this end, it examines its internal employee data. However, this would be likely to include some systematic biases. For example, the overwhelming majority of people who were hired and went on to successful careers in the 1970s, 1980s and 1990s were male and carried out military service. Moreover, not a single one of the successful project managers from this time period was certified as a Scrum master! Based on this data, it would be tempting to include military service as a relevant characteristic, while disregarding Scrum certification. However, any such model would certainly be ineffective, and would not reflect current social norms. This extreme example shows that input and training data may be correct, but still contain biases. Alternately, the data may simply be incorrect.

Errors can be detected and corrected at the level of individual data records. An individual right to **data disclosure and correction** can be helpful in this regard (see Chapter 4.1.1). This right is of benefit to the general public as well as to individuals. Citron and Pascale point to credit scoring as an example of this phenomenon:

> "*An important question is the extent to which the public should have access to the data sets and logic of predictive credit-scoring systems. We believe that each data subject should have access to all data pertaining to the data subject. Ideally, the logics of predictive scoring systems should be open to public inspection as well. There is little evidence that the inability to keep such systems secret would diminish innovation. The lenders who rely on such systems want to avoid default – that in itself is enough to incentivize the maintenance and improvement of such systems*" (Citron und Pasquale 2014: 26).

The authors writing from the Future of Life conference, which is focused on the development of public-welfare-oriented artificial-intelligence technologies, offer a rather more direct formulation:

> "*People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data*" (Future of Life Institute 2017).

These observers also go a step further, calling for the ability to correct personal data, as is stipulated in the German Federal Data Protection Act (BDSG) (§ 34 BDSG – Einzelnorm n.d.)

The general underlying idea here is simple: If everyone has the ability to see the data relating to them, and correct it if necessary, the issue also loses saliency among the general public. This conforms with the idea of so-called **data-methods solutions** – methods of data collection and evaluation that ensures data is current, accurate and complete. This is critical because this data forms the basis for the development of analytical and decision-making procedures. Control by individuals over the data relating to them offers a dependable means of ensuring the quality of the data corpus as a whole, and thus for increasing the overall reliability of algorithmic decision-making (Future of Privacy Forum 2017). However, it is unreasonable to situate responsibility for this task solely with the affected parties. Ordinary individuals can hardly be expected to review large, complex and frequently updated data sets to ensure that they are up to date, accurate and relevant. For this task, institutions are necessary.

The European General Data Protection Regulation (GDPR) also contains provisions on this topic, particularly in Chapter 3, articles 13 – 16, in the context of personal data (Europäisches Parlament und Rat der Europäischen Union 2016). These stipulate that upon request by an individual (here termed a "data subject"), entities responsible for data collection and processing must provide the data subject with a copy of his or her personal data (Art. 15 GDPR). In addition, every data subject has the right to make corrections to this data (Art. 16 GDPR).

Unanswered questions here relate primarily to the issue of implementation. In order to assert these rights, institutions are needed. In addition, experts and industry associations recommend the creation of a so-called **data ombudsperson or data-protection commissioner** (Otto 2017: 27).

Profile: Quality seal for the origin and quality of data
Key idea: Develop methods for ensuring that data is current, accurate and complete
Action area: Reviewing implementation
Stakeholders: Affected parties, system operators, researchers
Enforcing stakeholders: Policymakers, system operators, state
Instruments: Legislation (implementation), development of standards
Status: Ideas, early prototypes

In order to combat systematic biases in data (e.g., over- or underrepresentation of certain groups) that go beyond individual cases, instruments other than correction and complaint opportunities are needed. To this end, regulations for safeguarding the underlying data must be developed that collectively ensure the data is current, accurate and complete, while also taking the data's global usage into account. Understanding why an algorithmic system's output shows certain biases requires some knowledge of the type and origin of the training data. A helpful analogy here may be the **documentation of supply chains** in the food or clothing industries. Just as the ability to track the origin of processed chicken eggs all the way back to individual farms is important, data sets too – for example, those that have been used in the development of a facial-recognition model – must be traceable. To this end, standards are needed that ensure the provision of information regarding the **origin and type of the training database and data sets** (ACM 2017; FAT/ML 2016), for example with regard to representativeness, up-to-dateness, and so on. A research team has developed an initial proposal in this area – dubbed "datasheets for datasets" – and has implemented it in prototype form with some real training data sets (Gebru et al. 2018).

For the data-collection process, this means that whoever collects training data should tag it with standardized information regarding data type, foreseeable biases and limitations on use. A part of this solution could be mandatory minimum requirements for the labeling of certain data sets. Labeling of this kind would help improve the explainability and intelligibility of the decisions rendered by algorithmic systems:

> *"Develop standards to track the provenance, development and use of training data sets throughout their life cycle. This is necessary to better understand and monitor issues of bias and representational skews. In addition to developing better records for how a training data set was created and maintained, social*

> *scientists and measurement researchers within the AI bias research field should continue to examine existing training data sets, and work to understand potential blind spots and biases that may already be at work"* (Campolo et al. 2017: 2).

Another problem is the possibility of introducing distortions during data transfer. Distortions can occur when an ADM system interfaces with an informational database, or through the action of the software itself. Examples of this can be found in the highly sensitive sphere of DNA analysis within the law-enforcement context (Kirchner 2017). In this regard, comparing different systems provides a starting point for dealing with the problems.

A separate question is whether a regulation limiting the use of algorithms to the same context for which they were trained, or at least to comparable contexts, would be useful. For example, the use of medical diagnostic software developed with the help of a particular set of training data might be limited to populations corresponding to those contained in the training data. The crucial task here is to ensure the correspondence of aspects relevant to the software's goals. If the degree of this correspondence is not yet known, the software should not be used in a new, as yet insufficiently understood context (Otto 2018).

4.2.3 Legally enabling and ensuring the auditability of algorithmic systems

Profile: Reduce legal hurdles to auditability
Key idea: Legal restrictions on the review of algorithmic decision-making systems for research purposes must be eliminated or limited (e.g., based on catalogue criteria)
Action area: Reviewing implementation
Stakeholders: Policymakers
Enforcing stakeholders: The state
Instrument: Regulation
Status: Ideas

Civil, copyright and data-security laws should be changed to make an audit of algorithmic decision-making systems by external actors possible. Today, security researchers risk violating the tenets of criminal law if they try to examine an automated system's functions and security. Over the long term, restrictions of this kind would make abuses and failures difficult to trace, particularly in inclusion-relevant areas.

The starting point for all legal restrictions is the fact that algorithms are often protected by nondisclosure provisions. Training and usage data are generally under private control. In some circumstances, this could have serious societal consequences.

Following the Brexit referendum and Donald Trump's election as president of the United States, experts debated the degree to which automated perception-influencing technologies contributed to these outcomes. Underlying this question was the assumption that social networks' use of algorithms enhances the production of so-called **echo chambers**, in which individuals primarily come into contact with others who share their personal preferences. According to this theory, this leads over time to a polarization of society (although this trend was certainly already under way). At the same time, there was also debate over the influence of so-called **fake news** (false news reporting with manipulative intention), which, with the help of automated processes (social bots) and highly targeted advertisement technologies (micro-targeting), may have been able to shift the climate of public opinion.

Unfortunately, the social networks involved have not permitted any independent studies (Lischka and Stöcker 2017) on this issue. Thus, researchers have been unable to make a comprehensive study of algorithms' and automated technologies' impact on the online debates. The only study on supposed echo chambers in which the relevant user data could be directly evaluated was carried out by employees of Facebook itself. This study ostensibly shows that user preferences have a greater influence on message selection (news feed content) than do algorithms. However, even if algorithms did have a measurable influence on the selection and ranking of posts,

its significance remained unclear. The study had fundamental problems with regard to representativeness. It examined only individuals who were heavy Facebook users, and who had identified their political leanings there (a population totaling just 2 percent of users). Thus, the selection of users was not sufficiently shielded from interaction with the issue being examined. The study triggered broad debate regarding its methodology; however, it did not produce satisfactory research results (Sandvig 2015).

This example shows that the legal status quo in the area of algorithms and data can in some circumstances engender societal harm. A remedy here would be to carve out exceptions to trade-secret protections for the purposes of researching specific systems in individual cases, while also taking reasonable interests into account (Calo 2017) and adopting appropriate transparency provisions (Tutt 2016).

In addition to obliging system operators to provide relevant information, provisions could also be drafted making it easier to conduct **external research**. This relates largely to the use of automated data-collection methods on platforms such as the web (automated collection of publicly available information), or the establishment of multiple (automated) user accounts as a basis for **input-output analyses** (see Chapter 4.2.1). As a rule, these methods violate internet companies' general terms and conditions. At least until January 2018, such activities were regarded as unauthorized access to networked systems under **IT-security laws** such as the Computer Fraud and Abuse Act (1986), and were thus deemed violations of criminal law (Calo 2017; Stone et al. 2016). Currently, multiple pending lawsuits in the United States are seeking to eliminate legal uncertainties and protect research of this kind, as independent input-output analysis remains the most important method of detecting discrimination in algorithmic processes (ACLU 2017; Goodman 2015; Williams 2017). On 10 January 2018, the U.S. 9th District Court of Appeals issued a judgment that overturned previous case law, ruling that violating websites' general terms and conditions was not a criminal act (Williams 2018). Legislation in the European Union and in Germany should be scrutinized on this basis. For example, draft legislation currently under review by Germany's Bundestag proposes the introduction of a new criminal offense, **digital trespassing**, which like the Computer Fraud and Abuse Act would criminalize research of this kind (Buermeyer 2016; juris 2018).

Copyright laws, particularly the so-called **anti-circumvention provisions** of the Digital Millennium Copyright Act, are also an issue (ibid.). Potential legal violations here are similar to the violation of IT-security laws; here too, much of the focus is on automated data collection using automated user accounts or bots. For example, a disguised account (e.g., representing a "person with African American heritage") can be used to test the adequacy of facial-recognition systems, or as a means of investigating other similar issues. Numerous examples have shown that facial-recognition software often fails to recognize African Americans (AI Now Institute 2017), or falsely classifies them, for instance as gorillas (Doctorow 2018). For this reason, a review would seem to be extremely useful. However, the procedure – involving the use of automated accounts – requires the system operator to validate the accounts being used for the purposes of identity verification. This falsification for the purposes of research can violate the above-mentioned laws, as these also extend to technologies designed to restrict access to copy-protected works even if no copyright violations are committed. Local legislation should also be reviewed in these areas.

The current revision of the EU Database Directive, coming in the context of a broader EU copyright-law reform, also offers grounds for concern, as it denies legal certainty to numerous stakeholders that work with open data. This is relevant to algorithm auditing insofar as open-data databases can play a fundamental part in comparative algorithm testing and model creation (see Chapters 4.3.1 and 4.3.2).

Requiring system operators to produce documentation would also enhance the ability to audit algorithmic systems. Here, we refer to documentation that must be made available to the public or to regulators upon demand, and which focuses particularly on underlying models, algorithms and data (ACM 2017; Shneiderman 2016).

4.2.4 Institutionalizing the ability to object to algorithmic processes

Profile: Individual objection procedures

Key idea: Strengthen individual and collective abilities to object to ADM procedures, for example by creating a class-action right for consumer associations

Action area: Reviewing implementation

Stakeholders: ADM operators

Enforcing stakeholders: The state

Instruments: Regulation, institutions

Status: Ideas

Algorithmic decision-making systems pose considerable challenges from a due-process point of view, particularly with regard to the **contestability of decisions**. Conceivably, some ADM systems' decision-making processes will have to be documented in such a way as to support audits, in order to enable decision forensics work.

Addressing algorithmic decision-making procedures in Chapter 3, articles 21 - 22, the EU General Data Protection Regulation (GDPR) provides data subjects with **individual objection rights**, the **right to human intervention** by a data controller, and the right to **express his or her point of view** (Europäisches Parlament und Rat der Europäischen Union 2016). Researchers view these rights as being of critical importance (Citron 2008). Dreyer and Schultz (2018) provide a comprehensive assessment of the GDPR with a particular focus on algorithmic processes.

Profile: Class-action rights relating to algorithmic processes

Key idea: Strengthen individual and collective abilities to object to ADM procedures, for example by creating a class-action right for consumer associations

Action area: Reviewing implementation

Stakeholders: ADM operators

Enforcing stakeholders: The state

Instruments: Regulation, institutions

Status: Ideas

In general, objection rights must be institutionalized beyond the solely individual perspective. The possibility of establishing **class-action rights** should also be discussed. For consumer advocates, class-action rights have proved to be a strong mechanism for enforcing collective interests in the area of data protection. This is a promising tool particularly when rights associated with a group are violated (for example, through discrimination). Spindler's discussion regarding the expansion of class-action rights to encompass all data-protection laws and offenses against consumers in Germany is also applicable to algorithmic systems. Group related civil law legal remedies could supplement the public sector's legal oversight, he writes:

> *"In practice, the facilities for civil lawsuits currently granted to individual data subjects by the Federal Data Protection Act (BDSG), offering the prospect of damages, injunctive relief and so on, are not sufficient to enforce compliance with the data protection law at the civil law level. This is because individuals have insufficient incentive and lack the necessary information to implement this right on their own – even apart from problems of calculating damages, etc. Therefore, it is quite appropriate that individuals in such situations often see complaints to data-protection oversight agencies as the only way they can obtain help. However, this does not imply that this is the only entity that should take action; rather, the burden on supervisory authorities could be diminished through the creation of a class-action right"* (Spindler 2015: 2).

In an analysis of the GDPR, Roßnagel (2017: 38) employs similar logic in a call to bolster data-privacy protections through the creation of class-action rights. “The class-action right can be seen as an essential strengthening of data-privacy protections but will mean significantly more work for the oversight agencies.”

With regard to algorithmic systems, the roles played by **anti-discrimination associations** and the **German Federal Anti-discrimination Agency** (ADS) should be reviewed and expanded as appropriate. Today, neither has any class-action capabilities. This instrument can be useful if patterns of discrimination manifest more clearly and have a stronger impact at the collective level than at the individual level. In 2016, for example, an independent panel tasked with evaluating the German Non-Discrimination Act (AGG) proposed the creation of an **altruistic right of action** as well as a **limited class-action right** for the ADS:

> *“In cases of public interest, if no affected party is known that has been subject to specific disadvantages, the ADS should be granted an altruistic right of action that can be used to establish a breach of the AGG. This declaratory judgment should then have evidentiary effect in subsequent individual lawsuits”* (Berghahn et al. 2016: 191).

> *“Also conceivable would be a limited class-action right for the ADS that could be employed to allege breaches of the AGG – in agreement with those directly affected – without simultaneously asserting individual legal claims. The goal here would rather be the determination of the fundamental illegality of a behavior or a regulation”* (ibid.: 195).

4.2.5 Developing public oversight of algorithmic systems

The previous chapters focus on requirements and strategies for reviewing the functioning of algorithmic decision-making systems. However, we have for the most part skirted the question of who is, or who should be, responsible for a review or its elements.

Profile: Flesh out the need for specific oversight strategies and instruments
Key idea: Develop strategies distinguished by type of oversight (ex ante, ongoing, post facto), sector and overarching approach
Action area: Reviewing implementation
Stakeholders: Policymakers, system operators
Enforcing stakeholders: State, policymakers
Instruments: Institutions, approval procedures, seals of approval, certifications
Status: Ideas

A number of the strategies presented in this section imply that state oversight agencies should be created, or at least, that responsibilities should be defined and transferred to existing institutions. At its core, these actions require an identification of which algorithmic decision-making systems require oversight from an overall societal perspective, followed by the development of effective procedures that enable review of this kind. Whether implemented by regulatory authorities, an entity providing TÜV-like certification, or a ratings agency (Otto 2017), oversight and control options must be developed in cooperation with system operators. Proposals thus far have been insufficiently concrete. If the field is to be developed further, systematic analysis is needed that fully describes the status quo and answers the following questions:

1. Are there any substantial arguments against limiting the following points of analysis to algorithmic processes in use in Germany (our proposal for practicable restriction)?
2. In what sectors are algorithmic systems used today, or will they conceivably be used in the future?
3. In which of these sectors would the creation of a body to oversee algorithmic systems be useful?
4. What oversight institutions already exist in these sectors?
5. In reviewing the answers to questions 2, 3 and 4, where are gaps in oversight evident?

6. What forms of oversight (ex ante, ongoing, post facto) have already been institutionalized within the area of use?
7. What forms of oversight would be useful as applied to algorithmic systems?
8. In reviewing the answers to questions 6 and 7, where are gaps in oversight evident?

An empirical grounding (focused on areas particularly relevant to inclusion, if necessary) is critical if discussions regarding the need for algorithmic decision-making systems and appropriate strategies is to be productive. Moreover, this necessarily entails an examination of how existing oversight institutions in all relevant sectors (from aircraft accident investigation and drug approval to financial oversight and car registration) currently function. Are sector-specific solutions needed? Or should skills and competences be centralized (see also Chapter 4.4.2)? Where are algorithm approval procedures needed, and where would post-deployment review of an algorithm's use suffice?

Algorithm impact assessments (see Chapter 4.1.3) could form the basis for the initial evaluation of algorithmic systems. Information on an algorithmic decision-making system's goals and methods, data input quality, and expected results would facilitate its classification. This in turn could form the basis for allocating available resources and/or analytical capacities for the auditing of relevant cases**. A classification of the decision-making system** should take the following issues into account:

- The type and complexity of the algorithmic system
- The operator (e.g., public or private sector) and area of use
- Potential risks to affected individuals (Tutt 2016)

If the relevant analyses are made by state oversight agencies, audits within sensitive areas could be carried out under conditions of qualified transparency. Such analyses should examine not only the algorithmic systems themselves, but also relevant training data sets and the decision-making structures in which the systems are embedded.

4.2.6 Promoting civil society engagement

Profile: Establishing civil society watchdog organizations
Key idea: Civil society watchdogs expose novel problem areas related to automated decisions, and test possible solutions
Action areas: Reviewing implementation, reviewing optimization goals on the basis of compatibility with social norms
Stakeholders: Policymakers, civil society institutions
Enforcing stakeholders: State, non-governmental organizations
Instruments: Funding, regulation
Status: Early actors active today (e.g., AlgorithmWatch[16])

The state is not the only entity capable of exercising a watchdog function. In many areas, civil society actors have earned strong reputations for uncovering novel failures and problems with software systems. This is also true for algorithmic decision-making systems. In many cases, these entities have been the ones to make problems known to the public and to policymakers.

To give an example, the so-called **Heartbleed bug** in the OpenSSL software library demonstrated clearly that transparency alone is not enough to produce public visibility (Kroll et al. 2017: 10). The serious security flaws in OpenSSL, a standard software library used to encrypt internet data transmissions, had remained undiscovered for years despite the fact that the source code was generally available. After the flaw's discovery, the OpenSSL

[16] Transparency note: The Bertelsmann Stiftung (2017) funds the AlgorithmWatch organization ("Funding – More transparency and civil society overwatch of algorithms.").

Foundation revealed that maintenance and development of the software was essentially the task of a single full-time employee. The incident demonstrated that financial and staff resources are needed to ensure that transparency is translated into public awareness. This lesson led to the establishment of the **Core Infrastructure Initiative (CII)**, through which the Linux Foundation and large internet companies provide funding for developer positions and audits of the net's underlying software infrastructure (Diedrich 2014).

The algorithmic watchdog function does not relate exclusively to technical issues, such as those highlighted in the example above. Algorithmic decision-making system operators often treat factors like procedures and training data as closely held secrets, complicating attempts to analyse system processes externally. Yet, as the Australian Centrelink case (Rohde 2017) showed, civil society watchdogs can play a critical role even prior to a technical analysis. For this case, a piece of software had been developed in Australia that reviewed social benefit payments, with the goal of identifying individuals who had wrongly received unemployment benefits or social assistance. A computer program was tasked with reviewing cases and automatically correcting those that it found were suspicious. This process entailed automatically sending requests for repayment. In the course of its first analysis, however, the system rendered incorrect judgments in at least 20,000 instances. Affected individuals had to fight for months to appeal and correct these errors. Organizations representing affected individuals compiled cases of flawed individual decisions and publicized these errors in the media, thus kicking off a debate regarding the system's functioning.

Similarly, the debate over the prominent COMPAS system for assessing criminal offenders' risk of recidivism began only after ProPublica, a U.S.-based non-profit research group, invested considerable effort in researching, processing and evaluating the relevant data (Angwin et al. 2016).

These examples show that civil society oversight functions need resources not only for the technical analysis of systems, such as in the case of OpenSSL, but also for researching potentially problematic cases and actual usage practices, collecting data (by filing lawsuits, if necessary), developing data exchange standards, creating and maintaining test databases, and much more.

Role models in the environmental movement offer ideas for other potential areas of action. Here, non-governmental organizations function in part as ombuds bodies, but with low barriers to access. In the past, non-governmental organizations have similarly highlighted general problems by aggregating and evaluating individual cases, and then making public the results with the goal of triggering a societal debate. The establishment of civil society watchdog organizations focused on algorithmic decision-making systems is thus associated with calls for public responsibility on the part of system operators:

> *"After releasing an AI system, companies should continue to monitor its use across different contexts and communities. The methods and outcomes of monitoring should be defined through open, academically rigorous processes, and should be accountable to the public. Particularly in high-stakes decision-making contexts, the views and experiences of traditionally marginalized communities should be prioritized"* (Campolo et al. 2017: 1).

The experiences with cases such as COMPAS and Centrelink show that civil society watchdog organizations can play a critical oversight role. If this is to continue over the long term, however, appropriate organizations with relevant goals, proven competence, transparent and suitable methodologies, and relevant know-how must be identified and supported.

4.3 Achieving diversity

The issue of diversity within algorithmic systems is expressed on two different levels: 1) diversity of implementation and thus the various systems in use within a specific field of application (Chapter 4.3.1); and 2) diversity of

system goals, along with the spectrum of different operators in the public, private and civil sectors (Chapters 4.3.2 and 4.3.3).

Diversification on both levels is important for the following reasons:

- **Limiting a system's potential damage**: On both levels, system diversity (of goals and implementation) opens up the possibility of alternatives for affected individuals. A system error will affect people less profoundly if systems with different modes of functioning, and thus different effects, are also in place.
- **Innovation**: Competing approaches promote innovation.
- **Error correction**: Differences between the effects produced by different approaches and the degree to which goals are achieved can trigger new insights and encourage competition.
- **Social dynamics**: The diversification of algorithmic systems with different system goals, operating entities, and operationalization increases the likelihood that societal trends will be reflected more rapidly in at least some of the systems.
- **Comprehensive public welfare orientation**: Not every socially relevant goal can be meaningfully put into practice using a profit-oriented business model. It is therefore important to pursue a diversification of operating entities and optimization objectives: Public, private and civil society actors have different objectives, target groups and modes of functioning; they follow different principles and are occupied with different issues. Diversity within all these stakeholder groups helps ensure that the use of algorithmic systems promotes a broad-based common interest.

4.3.1 **Reinforcing diversification through accessible training** datasets

Profile: More accessible training datasets through public funding and regulation
Key idea: A prerequisite for diversification in algorithmic decision systems is the accessibility of relevant training datasets for the different providers
Action area: Diversity
Stakeholders: ADM developers, ADM operators
Public non-governmental organizations (research funding)
Legal support (research), institution-building
Status: Idea

In recent years, algorithmic systems have made enormous advances in image recognition, facilitated both by new hardware and the widespread availability of training data. In 2009, a team headed by computer scientist Fei-Fei Li published the ImageNet dataset, a database of 3.2 million indexed photos. This achievement was made possible by users of Amazon's "Mechanical Turk" crowdsourcing platform, who classified the contents of the images. The resulting dataset serves as ideal training material for weak artificial intelligence systems seeking patterns and/or correlations, as well as for the creation of models designed to recognize the content of new photos automatically (Deng et al., 2009). Fei-Fei Li has said that the success of his project in the annual ImageNet software competition was facilitated by a genuine paradigm shift, taking us closer to semi-automated learning on the basis of training data:

> *"The paradigm shift of the ImageNet thinking is that while a lot of people are paying attention to models, let's pay attention to data. Data will redefine how we think about models"* (Gershgorn 2017: 1).

Today, the dataset includes roughly 13 million photos, while the recognition rate of the winning software in the ImageNet competition has risen from 71.8% in 2010 to 97.3% in 2017 (l.c.). This demonstrates vividly the importance of training data. The methods first tested using the ImageNet dataset in 2010 were not completely new. Indeed, approaches making use of artificial neural networks had long been familiar within the field. What was new, however, was a dataset of this scope and quality that was easy accessibility.

Nonetheless, two important limitations of the ImageNet dataset should be noted: The quality of the results is not perfect, as indicated by the 97.3% performance rate. There are significant variations in accuracy across different skin colors, for example. In addition, the performance improvement has been facilitated by numerous other factors including improvements in camera technology and alternative forms of lighting.

The importance of large, proprietary samples is particularly apparent in the case of automated facial recognition. While publicly available data stores that serve a particularly important function as test samples do exist (e.g., a collection of 13,000 photographs of celebrities called "Labeled Faces in the Wild"), the best recognition rates by far for these samples are achieved by entities such as Facebook and Google who have the ability to use the many millions of user images uploaded to their proprietary systems to train their facial-recognition systems. Publication of these company-owned samples is unthinkable for privacy reasons.

Search engines and social networks represent further examples of valuable training data sources that are not freely accessible. This is one application of algorithmic decision-making encountered by a majority of internet users on a daily basis. The structuring, personalization and evaluation of content in social networks and search engines is performed by algorithmic systems that evaluate users' reactions as key signals (Lischka and Stöcker 2017: 15). No other provider can evaluate these reactions to develop their own recommendation systems. The concentration of users among just a few providers thus dramatically expands these entities' pools of proprietary data and gives them advantages over new competitors.

In 2014, Google and Facebook brokered more than half of all visits to online media – a share that has been rising for many years. Today, almost 75% of all traffic on the web can be traced back to posts or links provided by Google or Facebook (Staltz 2017). In 2015, the journal Science published a study on the usage behavior of Facebook users (Bakshy, Messing and Adamic 2015). The three authors were Facebook employees, and the data underlying their research is not available to anyone outside of Facebook.

A similar concentration is identified by Calo in the overall development of algorithmic systems:

> *"The reality that a handful of large entities (literally, fewer than a human has fingers) possess orders of magnitude more data than anyone else leads to a policy question around data parity. Smaller firms will have trouble entering and competing in the marketplace. Industry research labs will come to far outstrip public labs or universities, to the extent they do not already. Accordingly, cutting-edge AI practitioners will face even greater incentives to enter the private sphere, and ML applications will bend systematically toward the goals of profit-driven companies and not society at large. Companies will possess not only more and better information but a monopoly on its serious analysis"* (Calo 2017: 20).

In our view, the further development of the idea of open datasets must address three fundamental questions:

- In the balancing of interests, how can the legitimate business interests of commercial enterprises be taken into account?
- How can open training datasets and data privacy be reconciled?
- How should the topic of undesired subsequent reuse of open data be addressed?

One proposal for counteracting these tendencies toward concentration is to make any datasets produced through publicly funded research freely accessible to the general public. This would enable diverse algorithmic systems to use this "open data." The National Science and Technology Council takes this approach in a proposal included in a report to the U.S. president:

> *"Encouraging the sharing of AI datasets – especially for government-funded research – would likely stimulate innovative AI approaches and solutions. However, technologies are needed to ensure safe sharing*

> *of data, since data owners take on risk when sharing their data with the research community. Dataset development and sharing must also follow applicable laws and regulations, and be carried out in an ethical manner"* (National Science and Technology Council 2016: 31).

Public sector research funders and public welfare-oriented stakeholders should ensure that sufficient budgetary funds are made available for such measures. The British Royal Society recommends:

> *"Research funders should ensure that data handling, including the cost of preparing data and metadata, and associated costs, such as staff, is supported as a key part of research funding, and that researchers are actively encouraged across subject areas to apply for funds to cover this. Research funders should ensure that reviewers and panels assessing grants appreciate the value of such data management"* (Royal Society 2017: 8).

German political advisor Philipp Otto goes even farther, arguing that the state should go beyond simple research funding and take an active role in opening up access to data that is necessary to fulfill public-service duties, even if it stems from private sources:

> *"A discussion of the use of private-sector data in the public interest would involve many challenges, from data protection to ownership rights. Nevertheless, the significance of this issue extends beyond national borders. At the very least, a publicly controlled and/or publicly usable European data pool comprising data from public institutions and specific relevant data from the private sector, which is permitted to be used by the state under clearly defined conditions, promises to be an attractive thought experiment"* (Otto 2017: 30).

Any further development of the idea must address the question of whether datasets containing primarily personal information would be included. Is the robust anonymization of large training data sets practicable? And if so, how?

Data from the core areas of the state's general interest public service provision can be used to facilitate diversity within algorithmic systems, in large part simply by making this data available. Accordingly, the British foundation Nesta recommends:

> *"Certainly algorithms in fields like welfare to work, health or probation, that are paid for by taxpayers, should be as transparent as possible, and in particular training data should be open, since that's what – in many cases – shapes the algorithms"* (Mulgan 2016: 3).

Providing support for the collection and use of training data goes beyond the compiling and providing of access to such data in the form of a generally accessible training-data pool. Other conditions must also be met in order to increase algorithmic systems' explainability and auditability along with their diversity.

4.3.2 Using public sector demand for algorithmic systems to ensure diversity

Profile: The development by public sector entities of innovative algorithmic processes as a general interest public service

Key idea: Test and set exemplary standards for algorithmic systems as a general-interest public service, on a non-commercial basis

Action area: Diversity

Stakeholders: ADM developers, ADM operators

Enforcing stakeholders: The state

Instruments: Procurement, development, institutions

Status: Idea

The public sector can play an active role in the diversification of systems and operating entities, as demonstrated by the example of biobanks.. In the United States as well as Germany, various public institutions make use of algorithmic systems to prepare or make decisions (see Lischka and Klingel 2017; Rohde 2017). The standards for development, implementation and deployment demonstrated in this area do not, however, meet the requirements laid out within this paper for explainability and compatibility with social norms. There is a range of specific reasons for this, but numerous case studies highlight the following overarching deficiencies:

- State use of ADM systems demonstrates a lack of binding standards and processes (e.g., in terms of adequacy and explainability). How, for example, should processes be documented?
- State use of ADM systems demonstrates a lack of expertise in the design, implementation and evaluation of algorithmic systems.
- State use of ADM systems demonstrates a lack of ambition to develop exemplary solutions that go beyond the intended purpose, and which could serve as a model for other applications.

One of the expert commissions established by the Obama administration recommended that the state use its role in procuring, developing and using algorithmic systems to make a positive societal contribution. This recommendation included the following specific areas:

- **The promotion of open software and standards**: “To help support a continued high level of innovation in this area, the U.S. government can boost efforts in the development, support, and use of open AI technologies. Particularly beneficial would be open resources that use standardized or open formats and open standards for representing semantic information, including domain ontologies when available. Government may also encourage greater adoption of open AI resources by accelerating the use of open AI technologies within the government itself, and thus help to maintain a low barrier to entry for innovators. Whenever possible, government should contribute algorithms and software to open-source projects” (National Science and Technology Council 2016: 32).
- **The development of standards and procedures for use by the state:** “Agencies should work together to develop and share standards and best practices around the use of AI in government operations. Agencies should ensure that federal employee training programs include relevant AI opportunities” (Executive Office of the President et al. 2016: 16).
- Concentration of expertise with regard to **developing, implementing and evaluating algorithmic systems in a higher-level agency**: “The Federal Government should explore ways to improve the capacity of key agencies to apply AI to their mission“ (ibid.).

These various proposals are interlocking. Capacity-building regarding the development, implementation and evaluation of algorithmic systems depends on the extent to which these systems are intelligible, as well as on the expertise of those who design them. This, in turn, is made possible by open software and standards that form the basis for the informed development of standards and the exchange of best practices between public authorities.

Profile: The regulation of quality requirements for algorithmic systems in state software procurement

Key idea: Binding standards during public tenders for ADM systems as regards auditability, impact assessments, open standards

Action area: Diversity

Stakeholders: Government entities, ADM developers, ADM operators

Enforcing stakeholders: The state

Instrument: Procurement law

Status: Ideas, possible models in other areas

Social and environmental standards are often included in public-procurement award criteria. The development of, and adherence to, such standards is partly subsidized by the state, for example at the local level by the Germany-

wide network for fair public procurement, which is in turn is financed by the Federal Ministry for Economic Cooperation and Development. This instrument lends itself to the long-term furtherance of diversity in algorithmic decision-making systems, with regard both to the variety of systems and the variety of sectors.

Calo proposes:

> *"In addition, and sometimes less well recognized, the government can influence policy through what it decides to purchase. States are capable of exerting considerable market pressures. Thus, policymakers at all levels ought to be thinking about the qualities and characteristics of the AI-enabled products government will purchase and the companies that create them. Policymakers can also use contract to help ensure best practice around privacy, security, and other values. This can in turn move the entire market toward more responsible practice and benefit society overall"* (Calo 2017: 24).

The idea of using state-based purchasing power to promote public interest-oriented design behavior should help inspire procurement standards for algorithmic systems, along with the procedures established in other areas. Moreover, requirements regarding system explainability, diversity and compatibility with social norms, may facilitate the development of corresponding practices and tools. For example, in calls for tender for ADM systems, government entities could include binding instruments relating to stakeholder participation (see Chapter 4.1.4) or implementation assessment (see Chapter 4.2.1). Similarly, tender-granting institutions could establish requirements for the use and promotion of open standards and software, as well as for the availability of training data (see Chapter 4.3.1).

By using procurement procedures as leverage, public-sector entities have the option of requiring that the development and use of algorithmic decision-making systems be monitored. They can set minimum decision-forensics standards, ensure that training data is labeled and documented, and make sure that all these issues are included in impact assessments. A New York City legal amendment passed following the emergence of severe problems with the public administration's algorithmic decision-making systems offers one example of how such recommendations can be implemented. Among other provisions, the law calls for a group of experts to develop criteria for selecting any such systems to be used in the future (The New York City Council 2018).

4.3.3 Promoting the development of algorithmic processes in the public interest

Profile: Establishing funding programs and standards for public-interest-oriented development
Key idea: Support non-commercial, socially motivated ADM development
Action area: Diversity
Stakeholders: Researchers, ADM developers, ADM operators, non-governmental organizations
Enforcing stakeholders: state-funded non-governmental organizations
Instruments: Funding programs
Status: Idea; possible role models in other areas

Research and development funding is another instrument that can help push technology development in a public interest-oriented direction, especially through the use of state investments. Calo identifies a need to conduct more basic research and studies on the social embedding of algorithmic systems. Instruments discussed in Chapter 4.2, such as standardized impact assessments and codes of professional ethics, must be developed and tested. They offer practical starting points for applied research on how algorithmic systems are embedded within social contexts:

> *"Investment opportunities include not only basic AI research, which advance the state of computer science and help ensure the United States remains globally competitive, but also support of social scientific research into AI's impacts on society. Policymakers can be strategic about where funds are committed*

and emphasize, for example, projects with an interdisciplinary research agenda and a vision for the public good" (Calo 2017: 24).

As ideas on public interest-oriented research funding are further developed, the following aspects should be examined and fleshed out:

- Developing and implementing transparency and intelligibility standards for studies and projects, while making these a criterion for funding eligibility.
- Defining open-data and open-source standards and making these a criterion for funding eligibility.
- Making the transferability of research into practice a criterion for funding eligibility.

Public interest-oriented algorithmic system funding should not be focused solely on established research institutions. Projects conducted by volunteers within the open-source community or in the context of civil society initiatives also represent valuable sources of expertise (Chapter 4.2.6). The Prototype Fund is a good example of such activity. This public funding program, overseen by the Open Knowledge Foundation Deutschland and financed by the German Federal Ministry of Education and Research, supports "non-profit software projects in the areas of civic tech, data literacy and data security." The projects are supported through the initial demonstration stage of development, with both financial support and coaching offers available (Open Knowledge Foundation Deutschland n.d.).

The Prototype Fund is an example of diversity-directed funding in the start-up arena. In seeking to support diverse algorithmic-system operator models, state start-up funding can also focus more strongly on cooperative and non-profit forms of organization (such as non-profit limited liability companies), perhaps providing them with organizational and setup assistance as well as with financial resources. In the United States, the Wikimedia Foundation and the Mozilla Foundation have demonstrated that third-sector operator models can increase the range of diversity among algorithmic systems, even over the long term.

4.4 Creating favorable conditions for inclusion-promoting ADM system use

The challenges and possible strategies in the area of algorithmic decision-making are diverse. Thus far, we have outlined various options for state, private sector, research-oriented and civil society actors. In the following section, we will discuss overarching state tasks that can support and coordinate individual initiatives.

4.4.1 Reviewing legal frameworks for possible areas of adjustment

Profile: Analyzing the effectiveness of ADM regulations and their implementation

Key idea: Identify regulatory gaps, including in the implementation of existing regulations

Areas of activity: Assessing general framework conditions, reviewing goals, reviewing implementation

Stakeholders: Executive and legislative branches of government, ADM developers, ADM operators, non-governmental organizations

Enforcing stakeholders: State (regulation, analysis), researchers (analysis), non-governmental organizations (analysis)

Instruments: Legal analysis

Status: initial approaches

Two key questions must be answered with regard to the legal framework for algorithmic systems:

- Are there **regulatory gaps** in the legal framework?
- Are there **gaps in the way current laws are being implemented**?

As yet there has been no comprehensive, systematic analysis of these questions. As with the question of effective public oversight (see Chapter 4.2.5), there is a challenge here in bringing two aspects into harmony:

- The **specific characteristics of individual algorithmic systems** (see Chapter 3)
- The **specific characteristics of the individual** sectors in which algorithmic systems are being used

Jaume-Palas argues that the analysis should focus primarily on the usage environment:

> *"Algorithms exist in all areas (finance, health, business, education, transportation, industry, communication, etc. etc.). In some of these areas, oversight bodies and mechanisms already exist that are in a much better position to assess the context necessary for a judgment (for example, the German financial-market oversight body (the Federal Financial Supervisory Authority, BaFin), the German Federal Institute for Drugs and Medical Devices, or the technical inspection association (TÜV) for automobiles). An adaptation or readjustment of existing legislation may be required to enable these entities to carry out this function"* (Jaume-Palasí 2017: 1).

Any analysis of the legal framework will very likely conclude that there are a variety of gaps existing both in regulation and in the way these regulations are being implemented. Martini's analysis of Germany's General Equal Treatment Act (ACG), for example, offers some early indications of **regulatory gaps** in existing law as it relates to algorithmic systems:

> *"The regulatory objective of limiting the discrimination risk associated with algorithm-based procedures is closely aligned with the protective mission of the General Equal Treatment Act (ACG). Both are intended to prevent people at risk of discrimination – typically minorities – from being disadvantaged. To be sure, the ACG does not exclude software-based procedures from its area of application today; indeed, it is technology neutral in its conception. Nevertheless, it applies to only a limited set of areas – specifically employment, education, social services and other services available to the general public (Sections 2 and 19 AGG). With regard to agreements between private individuals, outside the context of labor relations the law applies only to so-called Massengeschäften (businesses that serve the general public without making distinctions between customers, such as hotels and supermarkets) and insurance companies – thus, entities ranging from dance clubs to health-insurance providers, no matter whether analogue or digital. In contrast, numerous specialized fields of application for software-based processes fall outside the ACG's coverage. Lex lata, it is worth considering an expansion of the catalogue of application scenarios contained in Sect. 2 Para. 1 of the AGG to a No. 9, covering unequal treatment between private entities that involves algorithm-based data assessment or an automated decision-making procedure"* (Martini 2017: 1021).

The creation of class-action rights for consumer-protection organizations, anti-discrimination associations and the German Federal Anti-discrimination Agency has been proposed, but such action lacks a foundation in current law (see Chapter 4.2.4). Thus, this represents a regulatory gap.

One particularly interesting form of regulatory gap relates to algorithmic procedures that should be banned**. Legal prohibitions** (if necessary, with waivers possible) could be based on algorithmic systems' structural features and/or sectoral characteristics. For example, some experts call for bans on the use of automated machine learning systems in certain inclusion-relevant areas (e.g., in legal, health care, education and social security contexts) if a minimum level of auditability and intelligibility (see Chapter 4.2) cannot be guaranteed for their individual outputs (Campolo et al. 2017; Eckersley, Gillula and Williams n.d.).

Other criteria for prohibitions could include security risks based on the field of use, or aspects of the system's specific construction (e.g., insufficient robustness). Similarly, fields of use and optimization goals that are at variance with overarching societal principles could also qualify. Where society has chosen a solidarity-driven

socialization of risk, as in the case of social insurance programs for example, algorithmic processes should not be allowed to re-individualize these specifically collectivized risks (Lischka and Klingel 2017: 7).

As a first indication of **gaps in the implementation** of existing law, we have already addressed the following points in this analysis:

- Methods for reviewing implementation (see Chapter 4.2.1)
- The legal framework for auditing algorithm use (see Chapter 4.2.3)

However, a number of fundamental questions regarding the **implementation of existing law** remain unanswered. For example:

- How can an algorithmic systems' optimization goals be effectively understood if they change rapidly and continually?
- How can these changes be analyzed on a retrospective basis (e.g., through detailed decision-forensics work using audit logs)?
- How can changes in the output of an algorithmic system be audited on a retrospective basis?
- How can causal links between changes in optimization goals, data inputs and system outputs be retrospectively established?
- To answer these questions, must standards and requirements for documenting optimization goals, implementation, outputs and inputs be created? Does this documentation need to include sufficient detail and be of sufficient quality to be used in legal contexts?

Martini addresses these implementation challenges specifically in the context of automated algorithmic machine-learning systems:

> *"For complex software applications in fields of use that involve the potential of individual harm, the necessity of ongoing oversight stems from the fact that such applications often change their behavior like a chameleon in the course of their operations – whether through updates, or because of machine-learning processes. A court ruling against a discriminatory software-based decision that comes into effect two years after the original legal violation is already long outdated"* (Martini 2017: 1021).

The example of **liability questions** illustrates the challenges that must be addressed with regard to **implementing existing law**. With a set of July 2017 changes to the Highway Traffic Act, Germany created a liability regime for autonomous vehicles that has yet to be tested. This stipulates that even if computers are being used, final responsibility lies with the humans. Thus, even if automated driving functions are engaged, the driver must remain alert and take control of the vehicle if the autopilot indicates it is necessary, or if the autopilot no longer appears to be providing adequate driving functionality. The new amendments thus confirm the principle of owner liability, with corresponding insurance obligations. In the case of accident, a kind of black box is supposed to be used to help clarify whether the incident was the result of technical or human failure. This device records essential data during the course of the vehicle's operation (Bundesregierung 2017; forum 2017).

In order to implement legal provisions of this kind, strategies are needed for **documenting an algorithmic system's various processes**, as well as the **interactions between its various components**. Initial proposals in this regard include the following:

- Require system operators to abide by **minimum standards of decision forensics** (Citron 2008: 1301 ff.; Tutt 2016).
- **Create insurance obligations** linked to the actual design of the algorithmic systems (ACM 2017; Shneiderman 2016).

4.4.2 Strengthening the state's regulatory capabilities

Profile: Developing a central agency for algorithmic systems, as well as sector-specific regulatory strategies
Key idea: Create an agency for algorithmic systems
Areas of activity: Assessing general framework conditions, reviewing goals, reviewing implementation
Stakeholders: The state
Enforcing stakeholders: The state
Instrument: Institutions
Status: Idea

The role of the state as a developer of, and contractor for, algorithmic systems used in the provision of public services was addressed in the context of ensuring diversity (see Chapter 4.3.2). State competences regarding algorithmic processes must, however, encompass the entire spectrum of use. This includes not only the commissioning of such systems, but also their assessment, oversight and regulation.

Reflecting the complexity of algorithmic decision-making in society, various authors propose developing state competences centrally or locating relevant competences within existing institutions. However, even given the presence of sector-specific characteristics, some overarching tasks should be located in a centralized **algorithmic-systems agency**. This is particularly true for the monitoring and analysis of the following points:

- Technological development and specific fields of use
- The assessment of cross-sectoral systemic risks
- The classification of algorithmic decision-making systems, for example on the basis of their mode of functioning, complexity, area of use and potential risk
- The identification of algorithmic decision-making systems that should be subject to regulation, generally meaning state monitoring or control in the form of auditing, certification or prohibition
- The exercise of oversight and control activities intended to ensure the proper use of data
- The development of security, research and use standards
- Regulation of liability and review procedures (Tutt 2016)

Most authors propose that a central institution of this kind should generally function as a consultative body that liaises with the legislature, the executive and the courts alike (Calo 2017; Cave 2017; Executive Office of the President et al. 2016; Mulgan 2016). However Tutt's proposal goes somewhat further, advocating for the creation of a central authority that carries out "soft-touch" regulatory tasks such as specifying transparency requirements and setting standards, but is also responsible for certifying and approving systems:

> "*The rise of increasingly complex algorithms calls for critical thought about how best to prevent, deter, and compensate for the harms that they cause. This paper argues that the criminal law and tort regulatory systems will prove no match for the difficult regulatory puzzles algorithms pose. Algorithmic regulation will require federal uniformity, expert judgment, political independence, and pre-market review to prevent – without stifling innovation – the introduction of unacceptably dangerous algorithms into the market"* (Tutt 2016: 1).

An **agency for algorithmic systems** with this orientation recalls the concept raised in 2016 by the German Federal Ministry for Economic Affairs and Energy regarding a federal agency for digital affairs with the goals of "bundling competences, supporting the political digital agenda, and sustainably developing digital literacy" (Bundesministerium für Wirtschaft und Energie 2016: 56). Similarly, the goals for an algorithmic-systems agency could be described as the establishment of legal, technical and societal competences, and their application to the public interest-oriented design of algorithmic systems.

On the other hand, according to Stone et al. (2016), there is also a need to develop this **expertise within existing institutions**, particularly in the area of state and public-sector tasks. Yet this requires the capacity to deal with algorithmic processes as a cross-sectoral issue. The goal here is to understand the interactions between algorithmic decision-making, political agendas and social objectives. Because algorithmic processes can only be judged in the context of their use, domain expertise is essential. Evaluations are context dependent. A review of algorithmic systems in the pharmaceutical sector, for example, would require knowledge of drug effects.

Developing competences inside authorities with sectoral experience offers the opportunity to align algorithmic decision-making systems more closely with political agendas. It allow offers the opportunity to incorporate specific administrative expertise into the design of algorithmic decision-making processes (Executive Office of the President et al. 2016; Stone et al. 2016).

State regulation has, as a rule, lagged behind technological development. This may in part be due to different issues of focus. State expertise is not limited to outcomes in a business context with relatively clear-cut optimization goals. Rather, it must keep the general welfare and all particular interests in view. By contrast, companies often hire experts from new fields and offer attractive career-advancement and remuneration opportunities (Schuetze 2018). Thus, the development of state competences both at the overarching level and in individual sectoral areas represents a considerable challenge. State-level activity offers the opportunity to preserve and expand sovereignty within the area of algorithmic decision-making systems, while at the same time, helping to shape the digital transformation (Calo 2017: 23).

4.4.3 Promote individual awareness and skills in dealing with algorithmic systems

Profile
Key idea: Improve potentially affected parties' ability to deal with algorithmic systems
Areas of activity: Assessing general framework conditions, reviewing goals, reviewing implementation
Stakeholders: Citizens
Enforcing stakeholders: State, businesses, non-governmental organizations
Instruments: Education concepts
Status: Initial ideas

Each of these strategies – stakeholder participation, objection procedures, information rights, and civil society watchdogs – requires citizen participation in order to be implemented. For example, those affected by algorithmic decisions must be able to access information, send details regarding their concerns to watchdog organizations, and assert their objection rights. This requires a certain level of basic knowledge regarding where algorithmic decision-making tools are being used, what opportunities and risks are associated with that use, and how a (potentially) affected individual can exert influence over the system's design and use. In this regard, responsibility cannot be situated solely with the individual.

Regardless of educational background or prior knowledge of algorithmic processes, every citizen must be in a position to defend her- or himself against questionable processes. This may entail recourse to skilled assistance from an institutional source. In this regard, the individual ability to act effectively must be accompanied by the strengthening and, if necessary, establishment of institutions that promote, support and supplement this individual-level capability and, where necessary, compensate for gaps.

However, even if a state has the capability to act effectively, and even if standards of professional ethics are in place, the general population also requires skills in dealing with algorithmic systems. The Royal Society defines these competences in the following way:

> *"(…) a basic grounding in what machine learning is, and what it does, will be necessary in order to grasp, at a basic level, how our data is being used, and what this means for the information presented to us by machine-learning systems"* (Royal Society 2017: 63).

A first step toward addressing this would be to gather information on the current state of public knowledge. It is currently unknown to what extent the general population is informed of how algorithmic systems are used and function. A first crucial initial task would be to develop an understanding of how competences in this area manifest and can be measured. In the United States, this concept has been discussed under the rubric of "algorithmic literacy," although it has not as yet been operationalized:

> *"This group also discussed algorithmic literacy, in terms of reading, writing and making algorithms. This solution is a response to a perception of growing 'illiteracy' and inequality in access to and control of algorithmic mechanisms. Algorithmic literacy programs are, in general, designed to enable more individuals to impact information flows and perceive when or if they or others are being marginalized"* (Caplan, Reed und Mateescu 2016: 8).

The question of developing skills within this area suggests an appeal to schools and universities – which the Royal Society (2017: 63) expresses as follows:

> *"If introduced at primary or secondary school, a basic understanding of key concepts in machine learning can help with navigating this world and encourage further uptake of data science subjects"* (Royal Society 2017: 63).

However, the question of how these skills can be strengthened for the majority of the population no longer in school remains unanswered. Possible actors in this regard include community colleges, consumer-protection organizations, data protection authorities and foundations. Civil society watchdogs could also develop information materials for the general public.

Lawyers Danielle Citron and Frank Pasquale offer a proposal that focuses on helping people develop such skills rather than on organizational forms, arguing that algorithmic systems' decision logic should be illustrated in the form of interactive models. They use the example of credit-scoring system simulations, which people can use to test how system output changes with different input values:

> *"Another approach would be to give consumers the chance to see what happens to their score with different hypothetical alterations of their credit histories.... To make it more concrete, picture a consumer who is facing a dilemma. She sees on her credit report that she has a bill that is 30 days overdue. She could secure a payday loan to pay the bill, but she'd face a usurious interest rate if she takes that option. She can probably earn enough money working overtime to pay the bill herself in 40 days. Software could give her an idea of the relative merits of either course. If her score dropped by 100 points when a bill went unpaid for a total of 60 days, she would be much more likely to opt for the payday loan than if a mere five points were deducted for that term of delinquency"* (Citron und Pasquale 2014: 29 f.).

However, this proposal currently lacks in concrete large-scale implementation of concepts. Initial questions to be answered in more detail include the following:

- What is the basis for the models used in the decision-making systems?
- How can system functioning be illustrated so that people without previous knowledge can use and learn from the interactive models effectively?

A few examples of such interactive approaches to communicating information of this kind do exist, such as the FICO credit-rating simulator (Free Credit Scores Estimator from myFICO n.d.), and the Justice.exe application developed by University of Utah computer-science students for a seminar on algorithmic decision-making (University of Utah Honors 2017). This latter piece of software puts the user in the role of a judge.

The actions of the accused are briefly described, along with their criminal histories and sociodemographic backgrounds. The user then decides whether to impose a maximum or minimum sentence. These decisions train a model, which after 50 completed cases displays the factors it has learned to recognize as strong signals for the maximum sentence, and which it can accordingly use to make its initial predictions. Was it the accused's criminal record? Skin color? Gender? Age? Marital status? Education? Within just five minutes, this program, based entirely on fictional data, demonstrates to users how machine-learning functions and what role human decision-making patterns play in the process. Regardless of what organizational framework is ultimately adopted to convey skills of this kind, this kind of software should certainly be developed, implemented and used.

Figure 4:

Source: Own illustration

5 Summary and conclusion: What next

Human actions and human decisions are fallible. People often discriminate unconsciously. In many situations, they have difficulty managing complexity, and make inconsistent decisions. **Algorithmic decision-making processes** can help in detecting and compensating for some of these failures, thus making decision-making processes more **consistent** and potentially fairer. This corrective function is a key argument for the use of algorithmic systems. For example, the unequal treatment of job applicants on the basis of characteristics society deems inappropriate for use as a decision basis (e.g., a foreign-sounding name) can be consistently prevented by algorithmic systems. Studies show that humans demonstrably tend to consider such inappropriate characteristics when choosing between job applicants.

In addition, algorithmic systems could improve the **quality** of analysis and decision-making. They can handle large amounts of data more **efficiently** than humans in some cases (faster, and possibly at a lower cost), and more **effectively** in others (producing greater overall benefit). For example, a search-engine's web crawler[17] works continually to analyze the relationships between billions of web sites. Software prepares radiological images for human analysis, for example by performing vertebral counts, thus making it possible to manage the explosive rise in image quantities (Harvey 2018). In hospital intensive-care units, software continually monitors changes in and interactions between multiple vital signs in all patients (Briseno 2018). In none of these examples could humans provide comparable performance under the same conditions without the support of algorithmic systems.

More evaluable data and new analytical procedures offer the opportunity to gain **new insights** both into individuals and into society at large. This can contribute to **qualitatively better decisions**, for instance through the consideration of more or better data in a given individual case. This in turn enables procedures to be tailored more closely to fit individual needs, while also allowing for better handling of complexity, for example in the allocation of resources.

However, if these opportunities are to be fully realized and be made to serve the public interest, the state, the private sector and civil society must all participate in shaping these developments. This is particularly important because, as we have seen, algorithmic decision-making systems too are fallible (see Chapter 2.2). Action is needed in four areas in particular:

- Ensuring that algorithmic systems' optimization goals are compatible with social norms
- Implementing these goals during the development and deployment phases (e.g., with regard to operationalizing goals, selecting the data to be used and embedding systems in their ultimate social contexts)
- Ensuring a diversity of systems and operator models in certain areas of use
- Creating favorable general framework conditions with regard to factors such as the legislative framework, as well as state and individual competences in dealing with algorithmic systems

In Chapter 4, this working paper offers an overview of potential strategies broken down by these four fields of activity (see Figure 4). Some of these approaches to shaping the field's development are briefly summarized below. The following aspects are of particular importance:

- **Legal background:** To what extent does the proposed strategy accord with current law? Is it covered by existing regulations? Is there further need for action? Would the implementation of relevant regulations pose any particular challenges?

[17] Computer programs for building and indexing web sites that automatically search and analyze the World Wide Web, and if required sort the results on the basis of certain criteria.

- **Implementation:** How concrete are the proposed strategies? Can they be easily implemented? Do the concepts propose situating powers or responsibilities within a specific institution, or do they entail the design of a new institution?

5.1 Goals and mechanisms: Assessing compatibility with social norms

It is impossible to establish optimization goals that are universally compatible with social norms across all application areas. Society, values and norms are in constant flux; algorithmic systems must reflect this dynamic instead of simply reproducing the past. Individuals that could be affected by such systems must be informed about them in order to be in a position to help influence their operation. Indeed, providing information for all participants, a task that can be carried out by various actors using a variety of procedures, is of paramount importance.

Developers, system operators, and stakeholders: Document interests and optimization goals

Anyone who develops or commissions algorithmic decision-making systems must document the various optimization goals selected and the balancing of interests this involves. All interests and stakeholder groups associated with the goals should be systematically identified, recorded, documented and included in the process.

By doing so, developers, system operators and users will make the values underlying their decisions visible. A suitable instrument in this regard is the development and documentation of an interest matrix depicting the variety of interests, stakeholders and possible optimization goals involved.

Developers, system operators, legislators: Provide affected parties with information about the use and goals of the ADM system

In order to be able to assess algorithmic decisions and exercise objection rights if necessary, affected parties must know the systems being used and have some insight into the systems' goals, methods and intended effects. To some extent, the EU General Data Protection Regulation (GDPR) specifies standards for this purpose in the realm of fully automated decisions. However, further regulation is also needed. To produce effective transparency, new procedures and instruments must be developed. Simply creating disclosure and transparency obligations is not enough; policymakers must also ensure obligations are carried out. The Chaos Computer Club's "**data letter**" proposal is one such concept. Under this model, system operators would be obliged to inform affected individuals regularly regarding which elements of their personal data have been stored, how this data has been linked to other information, and what profiles, assumptions regarding preferences, or evaluations (e.g., customer classes) have been derived as a result.

Application explanations (specifically, so-called counterfactual explanations) are another promising concept in this area. This approach is intended to help produce effective explanations of the algorithmic decisions made by machine-learning algorithms and complex systems. In the simplest case – credit approval, for example – an application explanation of this kind could provide information regarding how high an applicant's annual income would have to be for his or her rejected credit application to be approved. In the words of the concept's inventor:

> *"These counterfactual explanations describe the smallest change to the world that would obtain a desirable outcome, or to arrive at a 'close possible world'"* (Wachter, Mittelstadt and Russell 2017: 1).

At the same time, it must be ensured that the input parameters are correct – that is, that the underlying data is of high quality, free of bias and that an adequate decision logic is being used. Exerting oversight over algorithmic decision-making systems requires a detailed knowledge of their internal methodologies. This also makes great demands on consumer-protection organizations, which can exercise an important watchdog function.

Legislators, system operators: Reflect on and document expected outcomes and effects

Existing legal norms provide only limited answers to questions that extend behind the individual, for example regarding algorithmic decision-making systems' diversity or compatibility with social norms. Transparency vis-à-vis the general public regarding the goals, methods and results of algorithmic decision-making systems is just as critical as transparency vis-à-vis specific affected individuals. This is particularly true with respect to the systems' public acceptance, as well as in the creation of opportunities for members of the public to help shape the decision-making systems. A minimum level of knowledge regarding the use of the algorithmic procedures is necessary, however, if affected individuals or their representatives are to judge, critique or exert any influence over system operations.

Here, the so-called **"Beipackzettel"** (instruction leaflet) concept represents one initial suggestion of how system operators could be required to contribute to the goal of transparency. **Algorithm impact assessments** offer general information regarding the goals, methods and outcomes of algorithmic decision-making. In order to avoid discriminatory effects, these could be supplemented by information about competing systems in the same area of use. Impact assessments of this nature offer system operators the opportunity to choose between multiple systems and identify new sources of error (e.g., data distortions created at database interfaces). Lawmakers and system operators alike must work to develop new regulatory options. These can range from a self-regulated commitment to produce impact assessments to a legal obligation with the threat of penalties for those that fail to comply.

Developers, (institutional) users: Institutionalize methods of stakeholder participation

The only way to determine whether established optimization goals in fact conform with societal norms is if all relevant stakeholders are involved in the design of algorithmic systems. The development, implementation and use of algorithmic systems necessarily entails value-laden decisions from the start, for instance in the choice of data to be used, the methodology employed and the optimization goals chosen. Successful stakeholder-participation models from other areas have incorporated features such as the inclusion of trustees, review panels or entities representing affected parties. Models tailored for each specific institutional context must be developed and tested.

Developers, research and educational institutions: Establish codes of professional ethics

All actors participating in the design of algorithmic decision-making systems (that is, even beyond the mathematicians, engineers, computer scientists, and so on) bear a certain amount of responsibility. They implement predefined goals in code and processes, or otherwise make their contribution to the development of complex algorithmic decision-making systems. Members of this group in particular should reflect on their own roles and seek to sharpen their sense of social responsibility. A variety of proposals for establishing and implementing a code of **professional ethics** have been proffered. However, most of these tend to focus on the further development and concretization of binding standards, training objectives and advisory bodies. The objective is to codify process-based quality standards for the design of algorithmic systems in order to ensure adherence to certain minimum standards in the areas of appropriate diligence, explainability or impact assessment, for example.

5.2 Impact: Assessing the implementation of goals in algorithmic systems

Auditing algorithmic decision-making systems both on the basis of functionality and impact has been a key issue in the public debate. Moreover, it provides the foundation for the evaluation and further development of such systems in accordance with the identified goals.

Developers, researchers, regulators: Develop, test and institutionalize methods for auditing algorithmic systems

Traditional forms of **algorithm analysis (auditing)** are used to review many of the static systems in use today. The focus of such an audit is on the system's design and problem model, along with the implementation of its underlying logic. However, these procedures run into limitations when applied to complex, dynamic arithmetic systems. The integration of machine-learning algorithms in particular requires new forms of auditing that focus on reviewing inputs and outputs in accordance with knowledge of the underlying data. This currently highly dynamic field also includes the development of complex technologies for the **description of system operations**. Like application explanations, these are not suitable for reviewing input parameters, evaluating correct functioning or assessing compatibility with social norms. Rather, the focus here is on understanding how specific decisions are reached.

The development of standards for algorithmic decision-making systems is another topic closely associated with the creation of audit methods. This relates to procedural requirements, documentation strategies and encryption technologies.

Legislators, system operators, civil society watchdog groups: Prioritize audit tasks

Algorithm auditing is a resource intensive task. For this reason, **classifying algorithmic decision-making systems** in a way that helps convey the relative social necessity of various auditing tasks, and enables resources to be allocated on this basis, is a vital step. This prioritization should consider system type and complexity, area of use, and possible risks to affected individuals.

Legislators, (institutional) users, system operators: Institutionalize adequate auditing procedures

A variety of institutional strategies for assuring **effective oversight and transparency** are conceivable. In order to protect legitimate interests on the part of operators and/or data subjects (data-privacy protection), review panels, agencies or independent institutions could be established that guarantee a confidential review of algorithmic decision-making systems.

Legislators, system operators: Label the quality and origin of the data being used, create correction and usage standards

While reviewing entire algorithmic decision-making systems is vital, assessing their underlying foundation – the data being used – is also of critical importance. Ensuring that the data used in such systems is **correct, current and representative** is necessary.

The promise of machine-learning systems can be realized only on the basis of large data sets in their entirety. However, this data must be scrutinized in order to minimize risks such as intentional or unintentional discrimination. Biases in a data set can change the results of algorithms' learning processes.

The EU General Data Protection Regulation offers one strategy for protecting data quality, with its right to data access and correction giving individuals the opportunity to exert control over their own data. However, numerous exceptions and limitations make supplementary regulation appear necessary. For the right to be implemented in practice, public and private institutions will likely have to create data ombudsperson or data commissioner positions. In addition, some experts have proposed requiring data to be tagged with proof-of-origin information for use in the global data trade.

Legislators, civil society watchdog organizations, researchers: Institutionalize analysis of the data used in algorithmic systems

The proof-of-origin concept for data leads to a further topic: labeling data on the basis of its underlying nature, known biases or limitations on the way it can be used. We have seen that there are a number of possibilities for introducing biases or distortions in the course of data processing. In addition, a "functional comparability" between

training and post-deployment input data must be ensured, along with a certain level of transparency regarding data use in both the development and deployment phases.

Data sets should be certified as accurate, current and representative from the start; doing so creates opportunities for research and innovation. A lively exchange of best practices with regard to analyzing data sets could help produce significant pan-European progress in this area.[18]

Legislators: Adapt legal frameworks to facilitate algorithm auditing

Carrying out audits requires **access not only to the algorithms themselves**, but also to the **underlying data**. Facilitating this is a political challenge. Because the data underlying such systems is typically owned by private entities, such access is often impossible. In this regard, there is an urgent need to review the degree to which operators provide relevant information to state institutions, the public, civil society watchdogs and consumer-protection organizations, and to demand further transparency if necessary.

Legal reform is needed in order to facilitate different forms of external data access for the purposes of auditing. Examples from the research community have illustrated both the possibilities and limitations associated with automated methods ("web scraping") and collaborative research ("data donations"). While the former technique runs quickly against the limits established by system operators' general terms and conditions – in some circumstances enhanced by IT-security laws and copyright regulations – the latter approach has to date lacked both representativeness and sufficient speed. Some observers even regard forms of collaborative research as symptoms of a dysfunctional legal order. In this regard, applicable legal restrictions must be reviewed and eliminated if needed.

Legislators: Institutionalize public oversight of algorithmic systems

A number of the strategies presented here imply that state oversight agencies should be created, or at least that responsibilities should be defined and transferred to existing institutions. At its core, this requires an identification of which algorithmic decision-making systems require oversight from an overall societal perspective, followed by the development of procedures enabling such review. Whether this takes the form of an agency with regulatory approval powers, an algorithm-focused technical inspection association (TÜV) or a rating agency, oversight and control options must be developed in cooperation with system operators.

Legislators, institutional users: Create objection procedures and class-action rights

Closely linked with the audit of algorithmic decision-making systems for the purposes of evaluating compatibility with social norms is the institutionalization of **objection procedures or allowing human intervention** in order to assess whether a given decision is appropriate. This includes the right to:

- Obtain the intervention of an actual person for the purpose of reviewing a decision
- Present one's own point of view
- Appeal any algorithmic decision with serious and/or legal consequences

Realizing these rights requires that the algorithmic decision-making process be both explainable and transparent. Both of these aspects could be strengthened by **requiring system operators to document the data and decision methodologies being used**. As with all General Data Protection Regulation provisions cited here, the GDPR stipulations in this area contain numerous exceptions, and fall far short of covering all types of algorithmic decision-making systems. The need for further regulation on this topic must be assessed in each individual case.

[18] Telecommunications companies and top-level domain registries are promising partners in this regard, as they possess representative data on the internet usage of broad populations. With their help, biases deriving from incomplete or skewed data sets could be avoided. Differential-privacy procedures offer a promising means of avoiding privacy and data-protection risks in this regard.

In this regard, informal incentives and sanctions within the organizations or institutions using the ADM systems should be considered alongside formal legal provisions. More broadly, the international debate offers little in the way of concrete pointers regarding the implementation of objection procedures. In any case, it appears reasonable to consider the creation of **class-action rights**, which would allow public-welfare and consumer-protection organizations to aggregate and legally represent individual interests.

Legislators, civil society watchdog organizations: Develop algorithm-auditing resources

Effective funding for **civil society watchdog organizations** is needed in order to strengthen external research capabilities. Such groups have previously proven essential in the identification of new problem areas and solutions. However, human and technical resources are necessary in order to ensure that work on security holes and functional deficits can continue.

5.3 Diversity: Ensuring the diversity of algorithmic systems and processes

The development and use of algorithmic decision-making systems have to date been concentrated within the private sector, which controls the lion's share of necessary resources such as data and analytical procedures. As such, there is a risk of monopolization. In order to facilitate system diversity and foster public actors' long-term development and control capacities, data sets must be **made freely available, and support should be provided for their use**. These factors are equally as important as the **development of appropriate standards**. They offer an opportunity to counteract the information asymmetries developing between state and private-sector actors.

Legislators: Promote the public availability of data sets

Giving a variety of providers access to relevant training data sets is a necessary aspect of promoting diversity within algorithmic decision-making systems. Proposals here include:

- Providing public access to data derived from publicly funded research
- Providing public access to data from the public sector, particularly related to the provision of public services
- Providing public access to data from the private sector

Legislators, developers, researchers: Institutionalize public-interest-oriented data management

Existing proposals relate to:

- Developing business models based on publicly available data
- Promoting data-management mechanisms (data maintenance, storage capacities and requirements, etc.)
- Developing data standards (metadata)
- Developing minimum labeling standards (type and origin, biases, etc.)

Public sector: Using public-sector demand for algorithmic systems to ensure diversity

The public sector plays an active role in encouraging algorithmic decision-making system development that has a public-interest orientation. It does so by:

- Developing innovative algorithmic processes used to help public entities provide services
- Developing standards for the algorithmic decision-making systems used within the public sector

- Setting standards by making state purchases (contract criteria) – that is, through the mandatory inclusion of auditability, impact assessment and open-technology standards in public tenders for algorithmic systems
- Supporting open technologies that promote the public welfare

These propositions make it clear that the state's role in the area of algorithmic decision-making systems can in fact extend beyond that of a regulator. Indeed, it can act as an active partner shaping a positive climate of responsibility; thus, it should seek to do so without fail – if possible in cooperation with other actors.

5.4 Conditions: The law, state capabilities, individual competencies

The overview of challenges and potential strategies in the area of algorithmic decision-making makes it more than clear: State and individual competences alike must be fundamentally enhanced.

Legislators: Review the legal framework for potential gaps, and assess the implementation of existing laws

Generally speaking, existing legal codes must be examined for regulatory gaps such as the lack of class-action rights as well as for gaps in implementation. In this regard, both the characteristics of particular algorithmic decisions and the area or sector in which they are being applied must be considered. This also extends to legal prohibitions and questions of liability. Two key questions must be answered with regard to the legal framework for algorithmic systems:

- Are there **regulatory gaps in the legal framework?**
- Are there **gaps in the way current laws are being implemented?**

As of yet there has been no comprehensive, systematic analysis of these questions.

The EU General Data Protection Regulation establishes some initial requirements regarding the transparency of algorithmic decisions vis-à-vis affected parties ("data subjects"), at least in the area of fully automated decision-making systems. However, such tools are far less common than algorithmic decision support systems. Lawmakers must thus take further measures to secure transparency, by regulating the systems currently excluded from GDPR's provisions as well as algorithmic decisions that have no specifically legal impact on affected parties.

In addition, regulatory mechanisms should also be developed for the types of data that fall outside the GDPR's mandate. For example, this could include communications and transaction data, which constitutes the basis for new forms of profiling, as well as the use of anonymized data.

Legislators: Strengthen state capabilities

Overview capabilities regarding the **development, implementation and assessment** of algorithmic systems must be established. At a minimum, this must encompass:

- Technological development and specific fields of use
- The assessment of cross-sectoral systemic risks
- The classification of algorithmic decision-making systems, for example on the basis of their mode of functioning, complexity, area of use and potential risk
- The identification of algorithmic decision-making systems that should be subject to regulation, generally meaning state monitoring or control in the form of auditing, certification or prohibition
- The exercise of oversight and control activities intended to ensure the proper use of data
- The development of security, research and use standards

- Regulation of liability and review procedures (Tutt 2016)

These tasks could be exercised by an **agency for algorithmic systems**; such an entity could in turn have the overarching goals of developing and applying legal, technical and societal competences enabling it to promote the design of public-interest-oriented algorithmic systems. Thus, it would fulfill advisory, certification and approval functions.

These competences must also be developed within existing institutions, particularly where **state and public-sector tasks** are involved. The goal here is to understand the interactions between algorithmic decision-making, political agendas and social objectives. However, developing such competences also offers the possibility of aligning algorithmic decision-making systems more closely with political agendas, and of incorporating specific administrative expertise into the design of such processes (Executive Office of the President et al. 2016; Stone et al. 2016).

Legislators: Promote individual awareness and skills in dealing with algorithmic systems

Regardless of educational background or prior knowledge of algorithmic processes, every citizen must be in a position to defend her- or himself against questionable processes. The individual ability to act effectively in this area must be accompanied by the strengthening and if necessary the establishment of institutions that promote, support and supplement this competence and – if necessary – compensate for its lack. To this end, a survey of the population's state of knowledge is critical. In addition, training and continuing education courses must be created and supported; these can be offered by community colleges, consumer-protection organizations, data-protection authorities, foundations or civil society watchdog organizations.

5.5 Act now!

The proposals in this working paper focus primarily on the subject's technical and legal aspects. However, the summary highlights a number of integrative strategies in which state, private sector and civil society actors all assume various degrees of **responsibility**. Many constructive proposals still lack sufficient **specificity**, a concrete distribution of tasks or a concept for embedding these various aspects in **institutional structures**. Testing a number of different strategies in parallel would seem to be a promising way forward, as would coordinating related activities in this area.

As policymakers and other participants analyze this problem, they should give a higher priority to auditing existing strategies, including legal restrictions, than to the passage of over-hasty regulation. Overall, when considering regulation, it is important both to discuss the deficiencies in today's decision-making systems and to focus on the potential **opportunities** provided by **algorithmic decisions**.

Finally, it must be noted that this overview of potential strategies and areas requiring action in the context of algorithmic decision-making procedures has shown that **a wide variety of measures and methods offer opportunities for societal intervention and oversight, and indeed the ability to help shape the future of this field.** In no way do people seem to be at the mercy of machines. However, the opportunities and risks presented by individual cases must now be reviewed. Those who carry out this task must consider each system's area of use, complexity and degree of autonomy – and if necessary develop and test specific options for further action.

6 References

§ 34 BDSG – Einzelnorm (o. J.). https://www.gesetze-im-internet.de/bdsg_1990/__34.html (Download 29.3.2018).

Academic Advisory Council for Integration (2013). "'Soziale Teilhabe' Handlungsempfehlungen des Beirats der Integrationsbeauftragten." Federal Government Commissioner for Migration, Refugees and Integration. http://www.bagiv.de/pdf/soziale-teilhabe-empfehlungen-beirat.pdf (Download 22.4.2018).

American Civil Liberties Union (ACLU) (2017). "Sandvig v. Sessions – Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online." https://www.aclu.org/cases/sandvig-v-sessions-challenge-cfaa-prohibition-uncovering-racial-discrimination-online (Download 8.1.2018).

AI Now Institute (2017). "AI Now Public Symposium." https://www.youtube.com/watch?v=ORHe3dMvR2c (Download 22.4.2018).

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman and Dan Mané (o. J.). "Concrete Problems in AI Safety". https://arxiv.org/pdf/1606.06565.pdf (Download 22.4.2018).

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner (2016). "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Download 11.12.2016).

Association for Computing Machinery (ACM) (2017). "Statement on Algorithmic Transparency and Accountability." https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf (Download 22.4.2018).

Bakshy, Eytan, Solomon Messing and Lada A. Adamic (2015). "Exposure to ideologically diverse news and opinion on Facebook." *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160 (Download 22.4.2018).

Barocas, Solon, and Andrew D. Selbst (2014). "Big Data' s Disparate Impact." *California Law Review* (104) 3. 1–57. https://doi.org/10.15779/Z38BG31 (Download 22.4.2018).

Berghahn, Sabine, Vera Egenberger, Micha Klapp, Alexander Klose, Doris Liebscher, Linda Supik and Alexander Tischbirek (2016). *Evaluation des Gleichbehandlungsgesetzes.* Hrsg. Antidiskriminierungsstelle des Bundes. Berlin. (Also available at https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/AGG/AGG_Evaluation.pdf?__blob=publicationFile&v=15, Download 22.4.2018.)

Bertelsmann Stiftung (2011). *Soziale Gerechtigkeit in der OECD – Wo steht Deutschland? Sustainable Governance Indicators 2011*. Gütersloh. (Also available at http://news.sgi-network.org/uploads/tx_amsgistudies/SGI11_Social_Justice_DE.pdf, Download 22.4.2018.)

Bertelsmann Stiftung (2017). “Förderung – Mehr Transparenz und zivilgesellschaftliche Kontrolle von Algorithmen.” https://www.bertelsmann-stiftung.de/de/unsere-projekte/teilhabe-in-einer-digitalisierten-welt/projektnachrichten/foerderung-von-algorithmwatch/ (Download 12.4.2018).

Beuth, Patrick (2017). “Bombenbauer, die Aceton gekauft haben, kauften auch... “ *Zeit Online* 4.11. http://www.zeit.de/digital/internet/2017-11/amazon-terrorverdaechtiger-sprengstoff-zutaten-bestellt/komplettansicht?print (Download 22.4.2018).

Böttcher, Björn, Daniel Klemm and Carlo Velten (2017). *Machine Learning im Unternehmenseinsatz*. Hrsg. Crisp Research AG. Kassel. (Also available at https://www.unbelievable-machine.com/downloads/studie-machine-learning.pdf, Download 10.5.2018.)

Brennan Center for Justice (2017). “Brennan Center for Justice v. New York Police Department.” https://www.brennancenter.org/legal-work/brennan-center-justice-v-new-york-police-department (Download 4.1.2018).

Briseno, Cinthia (2018). “Wie Algorithmen Menschen vor einem frühzeitigen Tod bewahren können.” *Algorithmenethik* 15.3. https://algorithmenethik.de/2018/03/15/wie-algorithmen-menschen-vor-einem-fruehzeitigen-tod-bewahren-koennen/ (Download 13.4.2018).

Buermeyer, Ulf (2016). “’Digitaler Hausfriedensbruch’: IT-Strafrecht auf Abwegen.” *Legal Tribune Online* 6.10. https://www.lto.de/recht/hintergruende/h/entwurf-straftatbestand-digitaler-hausfriedensbruch-botnetze-internet/ (Download 22.4.2018).

Calo, Ryan (2017). “Artificial Intelligence Policy: A Roadmap.” *SSRN Scholarly Paper* No. ID 3015350). Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3015350 (Download 22.4.2018).

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker and Kate Crawford (2017). “AI Now 2017 Report (No. 2).” New York NY: AI Now Institute. https://ainowinstitute.org/AI_Now_2017_Report.pdf (Download 22.4.2018).

Caplan, Robyn, Laura Reed and Alexandra Mateescu (2016). “Who Controls the Public Sphere in an Era of Algorithms – Workshop Summary.” Gehalten auf der Who Controls the Public Sphere in an Era of Algorithms, Data & Society Research Institute. https://datasociety.net/pubs/ap/WorkshopNotes_PublicSphere_2016.pdf (Download 22.4.2018).

Cave, Stephen (2017). “Written evidence – Leverhulme Centre for the Future of Intelligence.” http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69702.html (Download 22.4.2018).

Christl, Wolfie (2014). *Kommerzielle digitale Überwachung im Alltag. Erfassung, Verknüpfung und Verwertung persönlicher Daten im Zeitalter von Big Data: Internationale Trends, Risiken und Herausforderungen anhand ausgewählter Problemfelder und Beispiele*. Vienna: Cracked Labs. (Also available at http://crackedlabs.org/dl/Studie_Digitale_Ueberwachung.pdf, Download 22.4.2018.)

Christl, Wolfie (2017). *How Companies Use Personal Data Against People – Automated disadvantage and personalized manipulation?* Working Paper by Cracked Labs. Vienna. (Also available at http://crackedlabs.org/dl/CrackedLabs_Christl_DataAgainstPeople.pdf, Download 22.4.2018.)

Citron, Danielle Keats (2008). "Technological Due Process." *Washington University Law Review* (85) 1. 1249.

Citron, Danielle Keats, and Frank Pasquale (2014). "The Scored Society: Due Process for Automated Predictions." *Washington Law Review*, (89) 1. 1–33.

City of Chicago (2017). "Strategic Subject List." https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np (Download 22.4.2018).

Cohen, I. Glenn, Ruben Amarasingham, Anand Shah, Bin Xie and Bernard Lo (2014). "The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care." *Health Affairs* (*33)* 7. 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048 (Download 22.4.2018).

Council of Europe – Committee of experts on internet intermediaries (2017). "Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications." https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a (Download 22.4.2018).

Dahllof, Staffan, Orr Hirschauge, Hagar Shezaf, Jennifer Baker und Nikolaj Nielsen (2017). "EU states copy Israel's ' predictive policing'." *EUobserver* 6.10. https://euobserver.com/justice/139277 (Download 22.4.2018).

Data & Society. Data Society Research Institute – Written evidence (AIC0221), Pub. L. No. AIC0221 and Select Committee on Artificial Intelligence (2017). "Response to UK House of Lords Call for Evidence." http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70517.html (Download22.4.2018).

Deng, Jia, Wei Dong, Richard Socher, Li-Ja Li, Kai Li and Li Fei-Fei (2009). "ImageNet: A Large-Scale Hierarchical Image Database." *Computer Vision and Pattern Recognition 2009.* CVPR 2009. IEEE Conference on. 248–255. http://www.image-net.org/papers/imagenet_cvpr09.pdf (Download 22.4.2018).

Dewes, Andreas (2018). "Begutachtung des Arbeitspapiers ‚Damit Maschinen den Menschen dienen'." (unpublished manuscript).

Diakopoulos, Nicholas (2016). “Accountability in Algorithmic Decision Making.” *Commun. ACM* (59) 2. 56–62. https://doi.org/10.1145/2844110 (Download 22.4.2018).

Dickey, Megan Rose (2016). “Police are increasingly using social media surveillance tools.” https://techcrunch.com/2016/09/23/police-are-increasingly-using-social-media-surveillance-tools/ (Download 22.4.2018).

Diedrich, Oliver (2014). “Linux Foundation finanziert OpenSSL-Entwickler.” https://www.heise.de/ho/meldung/Linux-Foundation-finanziert-OpenSSL-Entwickler-2213936.html (Download 10.11.2017).

Doctorow, Cory (2018). “Two years later, Google solves ‘racist algorithm’problem by purging ‘gorilla’ label from image classifier.” *boingboing* 11.1. https://boingboing.net/2018/01/11/gorilla-chimp-monkey-unpersone.html (Download 22.4.2018).

Dräger, Jörg, and Ralph Müller-Eiselt (2015). *Die digitale Bildungsrevolution. Der radikale Wandel des Lernens und wie wir ihn gestalten können*. München.

Dreyer, Stephan and Wolfgang Schulz (2018). What Exactly Does the General Data Protection Regulation Do with Regard to Algorithmic Decisions – and What Not? Ethics of Algorithms discussion paper #1: Bertelsmann Stiftung. Gütersloh.

Eckersley, Peter, Jeremy Gillula and Jamie Williams (o. J.). “Written evidence – Leverhulme Centre for the Future of Intelligence.” http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69702.html (Download 22.4.2018).

European Parliament and Council of the European Union (2016). “Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung), Pub. L. No. Verordnung (EU) 2016/679 (2016).” http://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32016R0679 (Download 22.4.2018).

Executive Office of the President, President’s Council of Advisors on Science and Technology (2014). “Big Data and Privacy: A technological Perspective.” https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf (Download 22.4.2018).

Executive Office of the President, National Science and Technology Council and Committee on Technology (2016). “Preparing for the Future of Artificial Intelligence.” https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (Download 22.4.2018).

Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) (2016). “Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.” http://www.fatml.org/resources/principles-for-accountable-algorithms (Download 1.2.2017).

Federal Government of Germany (2017). “Strassenverkehrsgesetz – Automatisiertes Fahren auf dem Weg.” https://www.bundesregierung.de/Content/DE/Artikel/2017/01/2017-01-25-automatisiertes-fahren.html (Download 16.1.2018).

Federal Office for Migration and Refugees (2017). “Moderne Technik in Asylverfahren.” 26.7. https://www.bamf.de/SharedDocs/Meldungen/DE/2017/20170726-am-vorstellung-modellprojekt-bamberg.html (Download 18.1.2018).

Federal Ministry for Economic Affairs and Energy (2016). *Digitale Strategie 2025*. Berlin. (Also available at http://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/digitale-strategie-2025.pdf, Download 22.4.2018.)

Ferguson, Andrew G. (2017). “Policing Predictive Policing.” *Washington University Law Review* (94) 5. 1113–1195.

Frick SJ, Eckhard (2018). “Welche Philosophie braucht die Medizin?” *Stimmen der Zeit* 2. 100–109.

forum (2017). “Autonomes Fahren ist nach Änderung des Straßenverkehrsgesetzes (StVG) erlaubt.” https://www.forum-verlag.com/themenwelten/kommunales/autonomes-fahren-ist-nach-aenderung-des-strassenverkehrsgesetzes-stvg-erlaubt (Download 16.1.2018).

frankro (2010). “Datenbrief.”

Free Credit Scores Estimator from myFICO (o. J.). http://www.myfico.com/fico-credit-score-range-estimator/ (Download 29.3.2018).

Future of Life Institute (2017). “Asilomar AI Principles.” https://futureoflife.org/ai-principles/ (Download 5.2.2017).

Future of Privacy Forum (2017). “Unfairness by Algorithm: Distilling the Harms of Automated Decision making.” Gehalten auf der RightsCon, Brüssel. https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision making/ (Download 22.4.2018).

Gallwitz, Florian (2017a). “Eine Polemik: Wie man mit einem würfelnden Schimpansen Terroristen fängt.” *Algorithmenethik* 21.12. https://algorithmenethik.de/2017/12/21/eine-polemik-wie-man-mit-einem-wuerfelnden-schimpansen-terroristen-faengt/ (Download 1.2.2018).

Gallwitz, Florian (2017b). “Begutachtung des Arbeitspapiers ‘Damit Maschinen den Menschen dienen’” (unveröffentlicht).

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford (2018). “Datasheets for Datasets.” http://jamiemorgenstern.com/papers/datasheet.pdf (Download 22.4.2018).

Georgieva, Petia (2017). “Written evidence – IEEE European Public Policy Initiative – Working Group on ICT.” http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69590.html (Download 22.4.2018).

German Bundestag 19. Legislative period (2018). “Schriftliche Fragen mit den in der Woche vom 29. Januar 2018 eingegangenen Antworten der Bundesregierung (No. Drucksache 19/605).”

German Medical Association (2015). “Musterberufsordnung für die in Deutschland tätigen Ärztinnen und Ärzte.” http://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/MBO/MBO_02.07.2015.pdf (Download 22.4.2018).

Gershgorn, Dave (2017). “The data that transformed AI research – and possibly the world.” https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/ (Download 5.11.2017).

Goodman, Bryce W. (2015). “A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection.” Gehalten auf der 29th Conference on Neural Information Processing Systems, Montreal (Kanada). http://www.mlandthelaw.org/papers/goodman1.pdf (Download 22.4.2018).

Gunning, David (2016). “Explainable Artificial Intelligence (XAI).” https://www.cc.gatech.edu/~alanwags/DLAI2016/%28Gunning%29%20IJCAI-16%20DLAI%20WS.pdf (Download 22.4.2018).

Harris, Elizabeth A. (2016). “Court Vacates Long Island Teacher’s Evaluation Tied to Test Scores.” *The New York Times* 10.5. https://www.nytimes.com/2016/05/11/nyregion/court-vacates-long-island-teachers-evaluation-tied-to-student-test-scores.html (Download 22.4.2018).

Harvey, Hugh (2018). “Why AI will not replace radiologists.” https://towardsdatascience.com/why-ai-will-not-replace-radiologists-c7736f2c7d80 (Download 7.3.2018).

Heaton, Brian (2015). “New York City Fights Fire with Data.” Government Technology 15.5. http://www.govtech.com/public-safety/New-York-City-Fights-Fire-with-Data.html (Download 22.4.2018).

Bundesgesetzblatt (1996). Internationaler Pakt über wirtschaftliche, soziale und kulturelle Rechte. (1966, Dezember 19). http://www.institut-fuer-menschenrechte.de/fileadmin/user_upload/PDF-Dateien/Pakte_Konventionen/ICESCR/icescr_de.pdf (Download 22.4.2018)

Jaume-Palasí, Lorena (2017). “Diskriminierung hängt nicht vom Medium ab.” *AlgorithmWatch* 3.7. https://algorithmwatch.org/de/diskriminierung-haengt-nicht-vom-medium-ab/ (Download 6.2.2018).

Jaume-Palasí, Lorena, and Matthias Spielkamp. (2017). “Ethics and algorithmic processes for decision making and decision support.” *AlgorithmWatch*. https://algorithmwatch.org/wp-content/uploads/2017/06/AlgorithmWatch_Working-Paper_No_2_Ethics_ADM.pdf (Download 10.5.2018).

Jedrzej, Niklas, Karolina Sztandar-Sztanderska and Katarzyna Szymielewicz (2015). *Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision-Making*. Warschau: Fundacja Panoptykon. (Also available at https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf, Download 22.4.2018.)

juris (2018). “Schutz vor digitalem Hausfriedensbruch.” https://www.juris.de/jportal/portal/t/1l1i/page/homerl.psml?nid=jnachr-JUNA180300598&cmsuri=%2Fjuris%2Fde%2Fnachrichten%2Fzeigenachricht.jsp (Download 22.4.2018).

Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi and Tom Blaser (2016). “Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making.” https://hbr.org/2016/10/noise (Download 25.3.2018).

Kirchner, Lauren (2017). “Putting Crime Scene DNA Analysis on Trial. *ProPoblica* 11.11. https://www.propublica.org/article/putting-crime-scene-dna-analysis-on-trial (Download 1.4.2018).

Kitchin, Rob (2016). “Thinking critically about and researching algorithms.” *Information, Communication & Society* (20) 1. 1–16.

Krafft, Tobias D., Michael Gamer, Marcel Laessing and Katharina Anna Zweig (2017). “1. Zwischenbericht Datenspende: Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017.” *AlgorithmWatch*. https://algorithmwatch.org/wp-content/uploads/2017/09/1_Zwischenbericht__final.pdf (Download 22.4.2018).

Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson and Harlan Yu (2017). “Accountable Algorithms.” *University of Pennsylvania Law Review* (165) 3. 633–705.

Kunichoff, Yana, and Patrick Sier (2017). “The Contradictions of Chicago Police’s Secretive List.” http://www.chicagomag.com/city-life/August-2017/Chicago-Police-Strategic-Subject-List/ (Download 22.4.2018).

Laskowski, Nicole (2017). “Machine learning’s training data is a security vulnerability.” *TechTarget* 31.10. http://searchcio.techtarget.com/news/450429272/Machine-learnings-training-data-is-a-security-vulnerability (Download 22.4.2018).

Lazer, David (2015). “The rise of the social algorithm.” *Science* (348) 6239. https://doi.org/10.1126/science.aab1422 (Download 22.4.2018).

Lenk, Klaus (2016). “Die neuen Instrumente der weltweiten digitalen Governance.” *Verwaltung und Management* (*22)* 5. 227–240.

Lischka, Konrad and Anita Klingel (2017). When Machines Judge People. Ethics of Algorithms discussion paper #1: Bertelsmann Stiftung. Gütersloh. https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/wenn-maschinen-menschen-bewerten/ (Download 22.4.2018).

Lischka, Konrad and Christian Stöcker (2017). The Digital Public. Wie algorithmische Prozesse den gesellschaftlichen Diskurs beeinflussen. Ethics of Algorithms discussion paper #1: Bertelsmann Stiftung. Gütersloh. https://doi.org/10.11586/2017028 (Download 22.4.2018).

Martini, Mario (2017). “Algorithmen als Herausforderung für die Rechtsordnung.” *JuristenZeitung* (72) 21. 1017–1026.

Mateescu, Alexandr, Douglas Brunton, Alex Rosenblat, Desmond Patton, Zachary Gold and Danah Boyd (2015). “Social Media Surveillance and Law Enforcement.” New York NY: Data & Society Research Institute. http://www.datacivilrights.org/pubs/2015-1027/Social_Media_Surveillance_and_Law_Enforcement.pdf (Download 22.4 2018).

Meyer, Thomas (2016). “Gleichheit – warum, von was und wie viel?” *Neue Gesellschaft/Frankfurter Hefte* 11. 42–46.

Mittelstadt, Brent (2016). “Auditing for Transparency in Content Personalization Systems.” *International Journal of Communication* (10) 0. 4991–5002.

Mittelstadt, Brent, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter and Luciano Floridi (2016). “The Ethics of Algorithms: Mapping the Debate.” http://philpapers.org/archive/MITTEO-12.pdf (Download 10.5.2018).

Moravec, H. (1998). When will computer hardware match the human brain. Journal of evolution and technology, 1(1), 10.

Mulgan, Geoff (2016). “A machine intelligence commission for the UK: how to grow informed public trust and maximise the positive impact of smart machines.” https://www.nesta.org.uk/sites/default/files/a_machine_intelligence_commission_for_the_uk_-_geoff_mulgan.pdf (Download 22.4.2018).

National Science and Technology Council (2016). “The National Artificial Intelligence Research and Development Strategic Plan.” https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf (Download 22.4.2018).

New York City Independent Budget Office (2016). “A Look at New York City’s Public High School Choice Process.” http://www.ibo.nyc.ny.us/iboreports/preferences-and-outcomes-a-look-at-new-york-citys-public-high-school-choice-process.pdf (Download 22.4.2018).

O'Neil, Cathy (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy* New York NY: Crown.

O'Neil, Cathy (2017). "Don't Grade Teachers With a Bad Algorithm." *Bloomberg.com* 15.5. https://www.bloomberg.com/view/articles/2017-05-15/don-t-grade-teachers-with-a-bad-algorithm (Download 22.4.2018).

Open Knowledge Foundation Deutschland (o. J.). "Prototype Fund." https://okfn.de/projekte/prototypefund/ (Download 3.11.2017).

Otto, Philipp (2017). "Leben im Datenraum – Handlungsaufruf für eine gesellschaftlich sinnvolle Nutzung von Big Data." *Perspektiven der digitalen Lebenswelt*. Eds. Herrmann Hill, Dieter Kugelmann and Mario Martini. Baden-Baden. 9–36. (Also available at https://irights-lab.de/wp-content/uploads/2017/06/Leben-im-Datenraum_Philipp-Otto_Perspektiven-der-digitalen-Lebenswelt_HillMartiniKugelmann.pdf, Download 22.4.2018).

Otto, Philipp (2018). "Begutachtung des Arbeitspapiers 'Damit Maschinen den Menschen dienen'" (unveröffentlicht).

Pasquale, Frank (2010). "Beyond innovation and competition: The need for qualified transparency in internet intermediaries." *Northwestern University Law Review* (104) 105. 105–171.

Pasquale, Frank (2016). *The Black Box Society: The Secret Algorithms That Control Money and information* (Reprint). Cambridge MA and London, England: Harvard University Press.

Passig, Kathrin (2017). "Fünfzig Jahre Black Box." *Merkur* (71) 823. 16–30.

Powles, Julia (2017). "New York City's Bold, Flawed Attempt to Make Algorithms Accountable." *The New Yorker* 21.12. https://www.newyorker.com/tech/elements/new-york-citys-bold-flawed-attempt-to-make-algorithms-accountable (Download 22.4.2018).

Prainsack, Barbara (2017). "Research for Personalized Medicine: Time for Solidarity." *Medicine and Law. World Association for Medical Law* (36) 1. 87–98.

Ramge, Thomas (2018). *Mensch und Maschine. Wie Künstliche Intelligenz und Roboter unser Leben verändern*. Stuttgart.

Rohde, Noëlle (2017). "In Australien prüft eine Software die Sozialbezüge – und erfindet Schulden für 20.000 Menschen." *Algorithmenethik* 25.10. https://algorithmenethik.de/2017/10/25/in-australien-prueft-eine-software-die-sozialbezuege-und-erfindet-schulden-fuer-20-000-menschen/ (Download 1.11.2017).

Roßnagel, Alexander (2017). "Zusätzlicher Arbeitsaufwand für die Aufsichtsbehörden der Länder durch die Datenschutz-Grundverordnung." https://www.datenschutzzentrum.de/uploads/dsgvo/2017-Rossnagel-Gutachten-Aufwand-Datenschutzbehoerden.pdf (Download 22.4.2018).

Royal Society (2017). *Machine learning: the power and promise of computers that learn by example*. London: The Royal Society. (Also available at https://royalsociety.org/~/media/policy/projects/machine-learning /publications/machine-learning -report.pdf, Download 22.4.2018.)

Russel, Stuart, Daniel Dewey and Max Tegmark (2015). "Research Priorities for Robust and Beneficial Artificial Intelligence. Association for the Advancement of Artificial Intelligence." https://futureoflife.org/data/documents/research_priorities.pdf?x56934 (Download 22.4.2018).

Russel, Stuart, and Peter Norvig (2012). *Künstliche Intelligenz. Ein moderner Ansatz.* 3., aktualisierte Auflage. Munich.

Sandvig, Christian (2015). "The Facebook 'It's Not Our Fault' Study." *Social Media Collective Research Blog* 7.5. https://socialmediacollective.org/2015/05/07/the-facebook-its-not-our-fault-study/ (Download 20.1.2017).

Sandvig, Ckristian, Kevin Hamilton, Karrie Karahalios and Cendric Langbort (2014). "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf (Download 22.4.2018).

Scherer, Matthew U. (2016). "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology* (29) 2. 354–400.

Scherer, Matthew U. (2017). "Public Risk Management for A.I.: The Path Forward." Gehalten auf der Beneficial AI 2017 – Asilomar Conference 2017. https://futureoflife.org/wp-content/uploads/2017/01/Matthew-Scherer.pdf?x56934 (Download 22.4.2018).

Schneider, Jan, Ruta Yemane and Martin Weinmann (2014). *Diskriminierung am Ausbildungsmarkt: Ausmaß, Ursachen und Handlungsperspektiven*. Saarbrücken. (Also available at http://www.svr-migration.de/wp-content/uploads/2014/11/SVR-FB_Diskriminierung-am-Ausbildungsmarkt.pdf, Download 22.4.2018.)

Schuetze, Julia (2018). *Warum dem Staat IT-Sicherheitsexpert:innen fehlen. Eine Analyse des IT-Sicherheitskräftemangels im Öffentlichen Dienst*. Eds. Stiftung Neue Verantwortung. Berlin. (Also available at https://www.stiftung-nv.de/sites/default/files/it-sicherheitsfachkraeftemangel.pdf, Download 22.4.2018.)

Selbst, Andrew D. (2016). *Disparate Impact in Big Data Policing* (SSRN Scholarly Paper No. ID 2819182). Rochester, NY: Social Science Research Network. (Also available at http://papers.ssrn.com/abstract=2819182, Download 22.4.2018.)

Shneiderman, Ben (2016). "Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight." *Proceedings of the National Academy of Sciences* (113) 48. 13538–13540. https://doi.org/10.1073/pnas.1618211113 (Download 22.4.2018).

Singer, Natasha (2015). “Bringing Big Data to the Fight Against Benefits Fraud.” *The New York Times* 20.12. https://www.nytimes.com/2015/02/22/technology/bringing-big-data-to-the-fight-against-benefits-fraud.html (Download 22.4.2018).

Spindler, Gerald (2015). “Stellungnahme zum Gesetz zur Verbesserung der zivilrechtlichen Durchsetzung von verbraucherschützenden Vorschriften des Datenschutzrechts – RegE BT-Drucks. 18/4631.” http://webarchiv.bundestag.de/cgi/show.php?fileToLoad=4246&id=1269 (Download 22.4.2018).

Stalder, Felix (2017). “Algorithmen, die wir brauchen.” *Netzpolirik.org* 15.1. 16. January 2017. https://netzpolitik.org/2017/algorithmen-die-wir-brauchen/ (Download 22.4.2018).

Staltz, André (2017). “The Web began dying in 2014, here’s how. https://staltz.com/the-web-began-dying-in-2014-heres-how.html (Download 5.11.2017).

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe and Astro Teller (2016). “Artificial Intelligence and Life in 2030.” Stanford CA: Stanford University. https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf (Download 22.4.2018).

Tene, Omer, and Jules Polonetsky (2017). “Taming the Golem: Challenges of Ethical Algorithmic Decision making.” *North Carolina Journal of Law & Technology* (19) 1. 125–173.

The New York City Council (2018). “A Local Law in relation to automated decision systems used by agencies, Pub. L. No. Int 1686-2017 (2018).” http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0 (Download 22.4.2018).

Torres, Phil (2017). “The Divide Between People Who Hate and Love Artificial Intelligence Is Not Real.” *Motherboard* 27.10. https://motherboard.vice.com/en_us/article/7x48kg/the-divide-between-people-who-hate-and-love-artificial-intelligence-is-not-real (Download 2.11.2017).

Tullis, Tracy (2014). “How Game Theory Helped Improve New York City’s High School Application Process.” *The New York Times* 5.12. https://www.nytimes.com/2014/12/07/nyregion/how-game-theory-helped-improve-new-york-city-high-school-application-process.html (Download 22.4.2018).

Tutt, Andrew (2016). “An FDA for Algorithms” (SSRN Scholarly Paper No. ID 2747994). Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2747994 (Download 22.4.2018).

University of Utah Honors (2017). “Justice.exe.” http://justiceexe.com/ (Download 22.4.2018).

Vieth, Kilian and Ben Wagner (2017). Calculated Participation. Ethics of Algorithms discussion paper #1: Bertelsmann Stiftung. Gütersloh. (Also available at https://doi.org/10.11586/2017025, Download 10.5.2018.)

Wachter, Sandra, Brent Mittelstadt and Chris Russell (2017). “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR” (SSRN Scholarly Paper No. ID 3063289). Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3063289 (Download 10.5.2018).

Web Foundation (2017). “Algorithmic Accountability. Applying the concept to different country contexts.” Washington DC: World Wide Web Foundation. http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf (Download 22.4.2018).

Williams, Jamie (2017). “EFF to Court: Accessing Publicly Available Information on the Internet Is Not a Crime.” *Electronic Frontier Foundation* 11.12. https://www.eff.org/deeplinks/2017/12/eff-court-accessing-publicly-available-information-internet-not-crime (Download 9.1.2018).

Williams, Jamie (2018). “Ninth Circuit Doubles Down: Violating a Website’s Terms of Service Is Not a Crime.” *Electronic Frontier Foundation* 10.1. https://www.eff.org/deeplinks/2018/01/ninth-circuit-doubles-down-violating-websites-terms-service-not-crime (Download 11.1.2018).

Zweig, Katharina Anna (2016). “2. - Working paper - Überprüfbarkeit von Algorithmen.” *AlgorithmWatch* 7.7. http://algorithmwatch.org/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/ (Download 7.9.2016).

Zweig, Katharina Anna (2018). Sources of Failure and Responsibilities in Algorithmic Decision-Making. Bertelsmann Stiftung. Gütersloh. (Also available at https://doi.org/10.11586/2018006, Download 10.5. 2018.)

7 About the authors

Julia Krüger is an independent social scientist from Berlin. She is interested in internet and digitization policy in international comparison, particularly with regard to the issues of regulatory content and manipulation, data and consumer protection, and algorithms and machine learning.

Konrad Lischka has written books, essays and blogs on digital society issues since 1999. After obtaining a journalism degree and training at the Deutsche Journalistenschule, he worked as editor-in-chief of bücher Magazin, and as deputy head of Spiegel Online's Netzwelt department. He subsequently switched to media and internet policy as a digital-society consultant for the North Rhine-Westphalia State Chancellery. From April 2016 until July 2018 he served as a project manager at the Bertelsmann Stiftung and as a project leader for the Ethics of Algorithms project.

Address | Contact

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telephone +49 5241 81-81322

Carla Hustedt
Project Manager
Ethic of Algorithms
Telephone +49 5241 81-81156
carla.hustedt@bertelsmann-stiftung.de

Ralph Müller-Eiselt
Senior Expert
Taskforce Digitization
Telephone +49 5241 81-81456
ralph.mueller-eiselt@bertelsmann-stiftung.de

www.bertelsmann-stiftung.de

BertelsmannStiftung