cooperation partner:

Working Group Monitoring of Algorithms | Translation: 13 January 2020
Original German version published on 18 March 2019

DENKIMPULS DIGITALE ETHIK:

Bias in algorithmic systems - explanations, examples and arguments

AUTHORS Corinna Balkow (Initiative D21 e.V.), Dr. Irina Eckardt (Initiative D21 e.V./ KPMG)

COLLABORATORS Johann Jakob Häußermann (Center for Responsible Research and Innovation Fraunhofer IAO), Lena-Sophie Müller (Initiative D21 e.V.), Ann Cathrin Riedel (LOAD e.V.), Dr. Nora Schultz (Office of German Ethics Council), Prof. Barbara Schwarze (Initiative D21 e.V./ Competence Center Technology-Diversity-Chancengleichheit e. V.), Birgit Wintermann (Bertelsmann Stiftung), Staff at KPMG AG and KPMG Law

- **People possess individual and unique socio-culturally shaped perceptions and experiences. This means that there can be no complex decisions that arise without biases – either in the analog or in the digital sphere.**
- **Algorithms do not act independently of the people who commission, produce or use them – thus duties and responsibilities arise for all people involved in the algorithmic system.**
- **There is no need for additional fundamental rights for the ethical handling of biases in algorithmic systems.**

I. Introduction

Scarcely any other development has changed our lives in recent years as much as the use of algorithmic systems. By relieving us of repetitive work and helping us make data-driven decisions, algorithmic systems make our lives easier. However, only very few people understand the basis on which these systems present their results to us, what data is processed to arrive at these results, and what happens to the data after it has been used.

This paper is based on discussions within the "Algorithmen-Monitoring" working group of the Initiative D21 e.V. The arguments presented aim to contribute to a more differentiated debate and to initiate a broader discussion in terms of how biases in algorithmic systems can be handled.

As part of the working process, pressing questions were identified and discussed from a socio-economic, technological and ethical-legal perspective, respectively. The talking points and illustrative examples refer to potential ways of coping with biases in algorithmic systems. It is not the aim of this paper to provide conclusive answers, but rather to present a basis for a more sustained approach to the issue.

What are biases?

Bias is commonly used to describe many things: from prejudices, distortions in data-driven decision-making, to the promotion or neglect of certain social groups.

The sources of biases can be both deliberate and unconscious.[1] They are based on individual experiences or lack of information on certain persons or groups. Data sets of marginalized ethnic groups, for instance, are only rarely selected for testing.[2] Similarly, not knowing that diseases can have different symptoms and effects depending on gender and/ or ethnicity, can result in poorly balanced and less representative data sets.[3] Both conscious and unconscious biases directly affect the quality of the results of algorithmic systems. Statistical quality standards from the "analog" world or DIN standards remain relevant measures for ensuring quality of processes in the digital realm. They are useful benchmarks for the transition to new digital standards. Nevertheless, scientific studies show that equality of opportunity in algorithmic systems cannot be achieved exclusively by mathematical methods, as there is often a requirement to interpret the exact same data differently depending on context.[4] A common effect associated with the presence of biases is the so-called biases blindness. It describes the tendency that most people consider themselves uninfluenced by biases. Studies in areas of management, for example, highlight that managers – no matter how qualified they are – cannot imagine how strongly they themselves are affected by biases.[5]

The world in which people operate – whether analog or digital – is very much influenced by subjective perception. This perception is based on individuals' social, cultural-historical and economic background, their socially constructed norms, education, but also by media and cultural institutions. These different influences are reflected in the discussions and decisions of individuals as well as of society. People often make decisions, despite having too little or too much information at hand. Many people overestimate the importance of the information they have or tend to only trust information that supports their previous opinion or knowledge. In addition, a decision, which was based on a certain result, is often based solely on the result itself, in complete disregard of the circumstances of how the result was reached. These limitations are referred to as **cognitive biases.**[6]

Due to the vast use of digital media and other technological solutions, more data is available today than ever before. People not only pass on their data, but with it they allow an insight to their views, opinions and assumptions. **Statistical biases** describe systematic or random errors in data collection as well as distortions in the distribution of data points. The existence of biases must be assumed in all data. These subsequently lead to erroneous or unwanted results in a statistical investigation. For example, the chosen design of a questionnaire can sufficiently influence the results: For a question of scale – „How do you estimate ... on a scale of 1-10 ?" – most respondents indicate a value in the middle such as 5 or choose a value at the extreme edges, i.e. 1 or 10.[7]

The necessary assumptions that must be made in the development of learning algorithms and during their actual implementation and realization, in order to be able to generalize observations, are called **inductive biases.** They form the basis of many algorithmic systems and therefore need to be considered specifically. Learning algorithms are based on data from the past, and thus do not automatically align with contemporary objectives: If in the past, men rather than women were employed or promoted, then the algorithmic system must be told whether this is still desired or whether it represents an undesirable distortion. However, generalization is impossible without inductive biases. In inductive learning, a function receives many individual examples and generalizes these step by step. The concept behind it is this: If a learning function can approximate a target function well by receiving a sufficiently large

[1] The Guardian (2018): Revealed: the stark evidence of everyday racial bias in Britain, online: https://www.theguardian.com/uk-news/2018/dec/02/revealed-the-stark-evidence-of-everyday-racial-bias-in-britain (accessed on: 02.02.2019).
[2] New Scientist (2018): Discriminating algorithms: 5 times AI showed prejudice, online: https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/ (accessed on: 02.02.2019).
[3] Deutsche Gesellschaft für geschlechtspezifische Medizin (2019): Sex in basic research: concepts in the cardiovascular field, online: https://www.dgesgm.de/images/pdf/Ventura-Clapier%20R%20Dworatzek%20E%20Seeland%20U%20et%20al%20Card%20Res%202017.pdf (accessed on 14.02.2019).
[4] A comparative study of fairness-enhancing interventions in machine learning (2018), online: https://arxiv.org/abs/1802.04422 (accessed on: 02.02.2019).
[5] Proov (2018): How Enterprises Overcome Digital Bias with International Collaboration, online: https://proov.io/blog/enterprises-overcome-digital-bias-international-collaboration/ (accessed on: 02.02.2019).
[6] Wikipedia (2018): List of cognitive biases, online: https://en.wikipedia.org/wiki/List_of_cognitive_biases (accessed on: 02.02.2019).
[7] Gesis (2015): Antworttendenzen in standarisierten Umfragen, online: https://www.gesis.org/fileadmin/upload/SDMwiki/Archiv/Antworttendenzen_Bogner_Landrock_11122014_1.0.pdf (accessed on: 02.02.2019).

set of examples, this function will also be able to approximate in the case of unknown examples.

Why do biases play a special role in algorithmic systems?

In simple algorithms (s.a. pocket calculators) biases will usually be disregarded. They should be in focus once people may be affected by an algorithm's automated calculations. For example, an internet search seems to many people like a simple, non-worrisome algorithm – you type in something, and the algorithm produces an output. Yet, in the background there is a complex data-based algorithmic system working. These are used worldwide, e.g. for online flight bookings[8], for application procedures[9] or in government organizations.[10] Reports on automated decision-making systems have repeatedly voiced criticism of discriminating algorithms.[11] In such cases, the biases in question are subjective in nature, whether intentional or unintentional. Some researchers distinguish between automated decision-making systems and supporting systems.[12] In this paper, the term "algorithmic systems" is used to describe the entire process from creation through to usage of a system as well as all people involved, and "algorithm" describes a unique set of instructions. For further discussion, it is important to first clarify the complexity of an algorithmic system.

The following illustration highlights the influence of biases throughout the entire life cycle of an exemplary algorithmic system. By considering different types of biases, better standards can be created that inform the identification and thus the suitable handling of biases. To help clarify the possible influences of different types of biases, the diagram outlines the different steps in the development of such a system. People will incorporate their prejudices regarding appearances or abilities of other people into their work on algorithmic systems as implicit value judgements. It is therefore important to consider the interests of the people who commission, plan, specify, develop, test and deploy algorithmic systems. In addition, the qualitative expertise of the data providers must be ensured, and social expectations and requirements must be considered.

[8] The Telegraph (2018): Airlines are starting to price their seats based on your personal information – but is it legal?, online: https://www.telegraph.co.uk/travel/news/dynamic-fare-pricing-airline-ticket-personalisation/ (accessed on: 02.02.2019).

[9] Independent (2018): Airlines face crack down on use of ´exploitative´ algorithm that splits up famlies on flights, online: https://www.independent.co.uk/travel/news-and-advice/airline-flights-pay-extra-to-sit-together-split-up-family-algorithm-minister-a8640771.html (accessed on: 02.02.2019).

[10] The Guardian (2018): ‚Dehumanising, impenetrable, frustrating': the grim reality of job hunting in the age of AI, available online under: https://www.theguardian.com/inequality/2018/mar/04/dehumanising-impenetrable-frustrating-the-grim-reality-of-job-hunting-in-the-age-of-ai (accessed on: 02.02.2019).

[11] Stats NZ (2018): Algorithm assessment report, online: https://www.data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf (accessed on: 02.02.2019).

[12] Gesellschaft für Informatik (2018). Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen. online http://www.svr-verbraucherfragen.de/wp-content/up loads/GI_Studie_Algorithmenregulierung.pdf (accessed on: 14.02.2019).

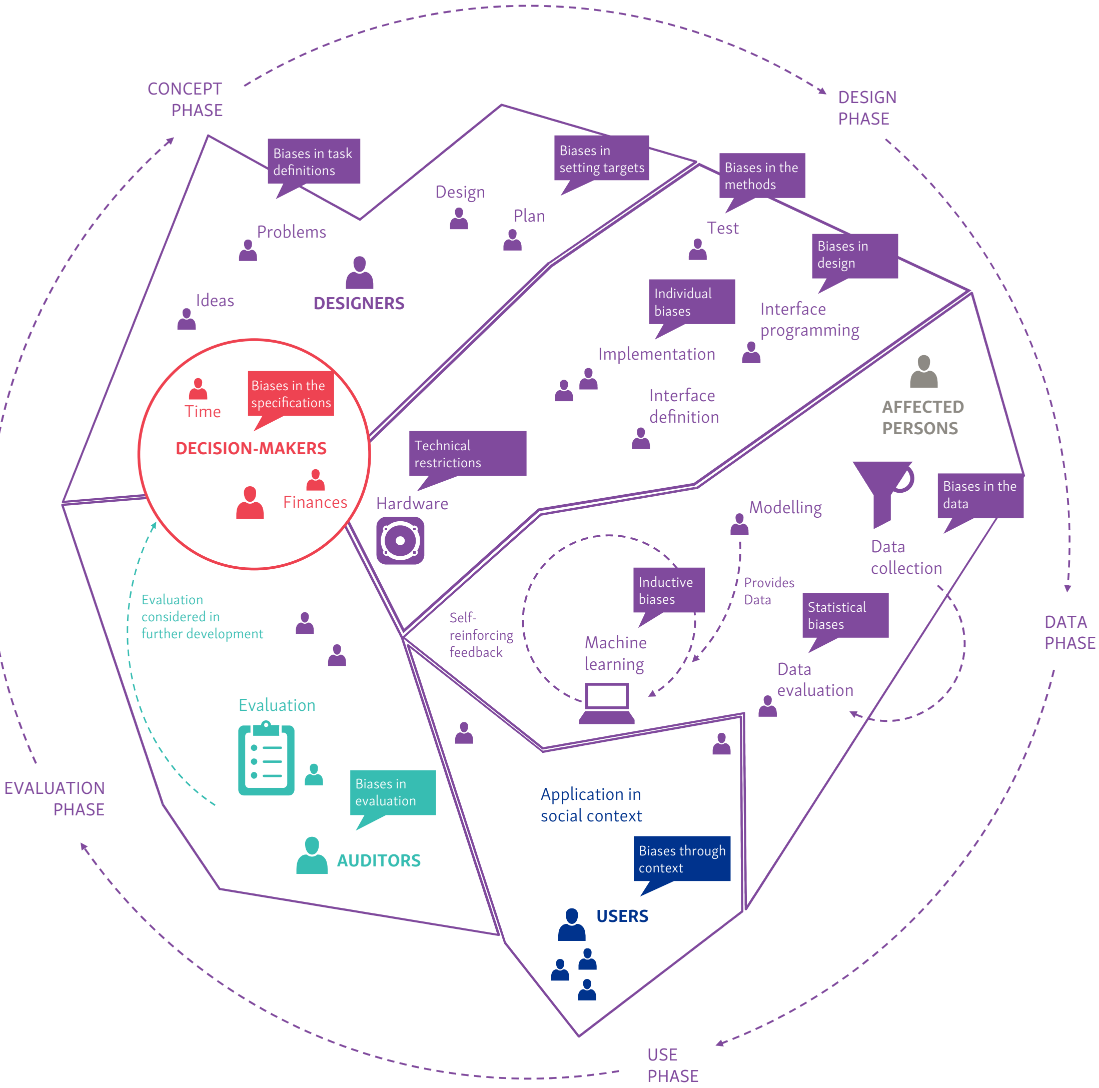

Figure 1: Locating potential biases in an algorithmic system

- **Concept phase – ideation of an algorithmic system:** Which problems should be solved by an algorithmic system? How is the objective defined? Which conditions regarding financing or time were given?
- **Design phase – implementation/ realization/ execution:** Which goals are defined for the algorithms? Which tasks can be implemented technically and how? Which options are programmed? Are test phases planned? What hardware is available?
- **Data phase – collection and organization of data:** Which data is perceived as relevant? Which data sets are prioritized? Is the available data suitable and sufficient for the goal of the algorithm? Is there a meaningful selection of training and test data? Are there statistical biases in the data?
- **Use phase – social embedding:** In which context is the algorithmic system used? Who uses the recommendations for their own decisions? Are the effects on social groups examined and tested?
- **Evaluation phase – area of assessment/ review and improvement:** How is success evaluated? What options exist to evaluate feedback? How is feedback taken into account? Are there ethically questionable outcomes?

II. Technological Perspective on Biases in Algorithmic Systems

The high complexity of algorithms associated with machine learning requires that the conceptual design of algorithm monitoring does not happen without the inclusion of technical experts. It is people who give an algorithmic system the necessary „world knowledge“ through their individual selection criteria. The system itself cannot make decisions about what is right or wrong, it can only provide probabilities. Each input must be integrated with its own data source. Each possible output needs to be predefined. The evaluation of the decisions based on outputs of algorithmic systems lie with the developers and the persons who use the algorithmic system.

If corrections are made to the algorithmic system, further checks must be carried out to ensure that the updated system functions accordingly. Since common procedures, applied technologies, integrated cooperation partners and used research and information sources frequently recur in companies and organizations, anti-biases research recommends the deliberate inclusion of heterogeneous perspectives.[13]

Argument: Data is biased

Description: Even with a careful selection of data sources, the data used will contain biases and these will thus auto-matically flow into the algorithmic system. It is the responsibility of the individuals who commission, develop or use algorithmic systems to investigate biases in the data and, if necessary, take appropriate measures to handle them.

Handling: Due to biases, some data sets are not suitable for their intended application in algorithmic systems. People who use algorithmic systems should identify biases in their data sets throughout the entire life cycle of an algorithm, test it against new or different data sources, and adapt or exchange data sources as appropriate. A regular review should become an integral part of the development and operational phase.

Examples: There are many technical possibilities available to identify biases and potential countermeasures. An open source approach can help to create transparency about which technical methods were used. For example, the University of Chicago presented an open source biases audit toolkit for machine learning developers, analysts, and policymakers; this can be used to examine machine learning models for discrimination and biases.[14] IBM presented solutions as part of the AIF360 project. These included examples for possible applications, types of handling biases, as well as concrete data-set examples.[15] Biases in data can lead to structural discrimination of groups, e.g. in the case of job-seeking for women as well as people with children.[16]

Argument: Biases are introduced by humans into the specific design of an algorithmic system.

Description: Through erroneous assumptions of people who commission, plan, design, implement, test and use the algorithmic system, they both consciously and unconsciously bring biases into their decisions in respect of the aim of the system, the selection of data sets, the social contextualization, and the expected results. In their roles, they have the task not only of assessing the functionality of a system, but also to question the existence of biases and to compare the existence with known typical errors.

Handling: In order to enable humans to identify existing biases, it is first necessary to enable humans to recognize biases. One way to do this is via initiatives that focus on greater diversity around algorithmic systems: Black in AI[17], Queer in AI[18], Women in Machine Learning[19], Lesbians Who

[13] prooV (2018): How Enterprises Overcome Digital Bias with International Collaboration, available online under: https://proov.io/blog/enterprises-overcome-digital-bias-international-collaboration/ (accessed on: 02.02.2019).

[14] Center for Data Science and Public Policy of The University of Chicago (2019): https://dsapp.uchicago.edu/projects/aequitas/ (accessed on: 14.02.2019).

[15] Detecting and mitigating age bias on credit decision (2019), online: https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb (accessed on: 02.02.2019).

[16] Futurezone (2018): Computer sagt nein: Algorithmus gibt Frauen weniger Chancen beim AMS, online: https://futurezone.at/netzpolitik/computer-sagt-nein-algorithmus-gibt-frauen-weniger-chancen-beim-ams/400345297 (accessed on: 02.02.2019).

[17] Black in AI (2019): https://blackinai.github.io/#about (accessed on: 02.02.2019).

[18] Queer in AI (2019): https://queerai.github.io/QueerInAI/ (accessed on: 02.02.2019).

[19] Women in Machine Learning (2019): https://wimlworkshop.org/ (accessed on: 02.02.2019).

Tech[20], Latinx in AI[21], Speakabled[22] or AI for Deaf[23]. It is essential not only to restrict oneself to existing initiatives, but also to examine whether other groups of people should be considered, as each context of application provides unique social dynamics and thus unique requirements.[24]

Examples: When designing algorithmic systems, many possible erroneous assumptions may arise, such as: "a product has exactly one price", "there are exactly two sexes", or "names of men do not change over time". A collection of such erroneous assumptions[25] can play an essential role in identifying and avoiding biases. The use of non-comparable data sets can also lead to biases that jeopardize the desired outcome of the algorithmic system. An example can be found in the current discussion surrounding a program which judges whether or not you are criminal from your facial features.[26]

Argument: People often find it difficult to understand how the algorithmic system works and are thus either unable to recognize biases or recognize them only to a very limited extent.

Description: Users and civil society alike require transparency from the owners of algorithmic systems. Only through transparency can these persons understand how, why and which biases potentially exist as well as for which tasks the algorithmic system is used. Due to the confidentiality of the data or the proprietary use of the data, in many cases this cannot be done directly. In such cases, an external verification of biases in algorithms and their underlying data can take place by third parties.

Handling: Creating transparency and comprehensibility about potential biases and the handling of biases in algorithmic systems is based on the results of technical analyses. These should be made public by the developers and owners. In many cases, where a direct disclosure of the data or the algorithm is not possible or not desired, analyses can be carried out by third parties, with subsequent publishing of results. This facilitates transparent analyses without full disclosure of individual data sets or algorithmic codes. However, it is imperative to offer transparency over what was examined during the audit and which methods were used.

Examples: When undergoing a credit assessment, people are treated as objects by the algorithmic system. There are efforts on the part of public organizations[27] to create transparency about algorithmic systems and biases. One project, for example, attempts to understand the assumptions and modes of action taken by an algorithm measuring the creditworthiness of people.[28] Even the active use of an algorithmic system can show negative biases effects. For example, people can usually only agree to a general use of their data in exchange for a service free of charge. Rarely can they limit the use of their data, such as location, duration, or frequency. This renders many helpless and unable to develop digital intuition.[29]

[20] Community of Queer Women in or around tech (2019) https://lesbianswhotech.org/about/ (accessed on: 02.02.2019).
[21] LatinX in AI (2019): https://www.latinxinai.org (accessed on: 02.02.2019).
[22] Liste von Menschen mit Behinderungen, die über Tech und Programmierung sprechen können. (2019): https://www.speakabled.com/sprecherinnen/?members_search=Tech+und+Programmierung (accessed on: 02.02.2019).
[23] Rochester Institute of Technology (2019): https://www.ntid.rit.edu/ (accessed on: 02.02.2019).
[24] BBC News (2018): IBM launches tool aimed at detecting AI bias, online: https://www.bbc.com/news/technology-45561955 (accessed on: 02.02.2019).
[25] Awesome falsehood (2019): https://github.com/kdeldycke/awesome-falsehood/blob/master/README.md (accessed on: 02.02.2019).
[26] Motherboard (2016): A New Program Judges If You're a Criminal From Your Facial Features, online: https://motherboard.vice.com/ en_us/article/d7ykmw/new-program-decides-criminality-from-facial-features (accessed on: 02.02.2019).
[27] Open Knowledge Foundation (2019): https://okfn.de/ (accessed on 14.02.2019), Algorithm Watch (2019): https://algorithmwatch.org/de/mission-statement/ (accessed on: 02.02.2019).
[28] Open Knowledge Foundation (2019): Get Involved: We crack the Schufa!: https://okfn.de/blog/2018/02/openschufa-english/ (accessed on: 02.02.2019).
[29] Müller, Lena-Sophie (2016) Das digitale Bauchgefühl. In: Friedrichsen M., Bisa PJ. (Hrsg.) Digitale Souveränität. Springer VS, Wiesbaden.

III. Social-Economic Perspective on Biases in Algorithmic Systems

By allowing algorithmic systems to make decisions where previously humans did this, society is given the chance to correct biases. System outcomes and processes can be critically questioned and compared with human decision-making.

Argument: Due to increasing confrontation with biases in algorithmic systems, „analog“ biases will increasingly be put to the test.

Description: Human decisions are often based on socially and culturally constructed norms that contain stereotypes and thus, also biases. These analog biases are rarely questioned or criticized as long as other social groups benefit from adhering to them.[30]

When algorithmic systems are introduced, people make higher demands and are more critical about biases. One of the reasons for this is that instructions in the analog world are re-defined with a certain margin of discretion and this margin of discretion is usually granted to some people. In the digital world, on the other hand, these instructions seem fixed and clearly defined.[31]

Handling: By educating all involved actors as well as the wider society about biases in algorithmic systems, they can develop their own thoughts and opinions on the subject. The identification of potential biases in digital systems requires a review of previously analog practice. Such education can be conveyed through digital and analog media, in the form of lectures, workshops or training, but also through public-opinion-forming journalism.

Companies, science- and research institutions, as well as public and private institutions can work together to make use of each other's knowledge. By exchanging information, they can independently grow and evolve in a positive manner. In the future, they will be able to better identify and regulate biases in algorithmic systems as well as in the analog world.

Examples: The far-reaching effects of biases can be illustrated through the example of a bookstore. When asking for a book recommendation in a local bookstore, the preferences of the employees, the current bestseller list and the selection of books in the store play a role in the recommedation. It is widely accepted that the recommendations of professionals are based on their experience. But if an algorithmic system is trained with these experiences and increases the reach, the decisions taken based on an individual professional's experience have completely different, vast implications.
Similar to this example, it is also widely accepted that students are assigned to specific schools or classes according to different, subjective selection criteria. The basis for these decisions are often not made public and remain largely discretionary. If this were to be automated and instead be executed by an algorithmic system, the set of criteria that inform the systems decisions on school placement are no longer subjective, but clearly defined. If this were also to be desired by the municipalities and schools themselves, there would be more chances for transparency in terms of which criteria determine which school a child attends.

Argument: No decisions exist without biases.

Description: Socio-cultural experiences, learning outcomes at school as well as economically conditioned living conditions influence how people take in and process information. These individual interpretations of information form the basis of our decision-making. The aim should therefore be to, reduce biases by being conscious and aware of their existence. Already existing quality standards from the analog and digital world can be used and adapted to reduce biases in algo-rithmic systems. For example, professional ethics-, quality- and anti-discrimination standards, user integration- and legal equal-opportunities-requirements can be translated and integrated into a "digital requirements catalog". This catalog can be used to inform ethical handling of biases for each of the respective phases of the development process. Consequently, there are various possibilities to reflect on
biases and to make biases in analog and algorithmic

[30] Hohlweg, Jelena; Salentin, Kurt (2014): Datenhandbuch ZuGleich. Zugehörigkeit & (Un-) Gleichwertigkeit IKG Technical Report Nr. 5, Version 1. Bielefeld, online:

[31] Magazin Mitbestimmung (2018): Frau Zweig, was können Computer besser, und was Menschen?, online: https://www.magazin-mitbestimmung.de/artikel/Frau+Zweig%2C+was+k%C3%B6nnen+Computer+besser%2C+und+was+Menschen%3F@6032 (accessed on: 02.02.2019).

systems more transparent.

Handling: Before moving to action, one must first reflect and identify whether unintended biases are present in each algorithmic system. Based on specified quality criteria[32], an ability to judge or evaluate existing algorithmic systems can be created. When most biases are identified, it must be decided whether and how to proceed with the system. As a supplement to basic standards, checklists for higher standards could be developed.

Possible points on a checklist could be:

- How do you create a distinctive understanding in your company of the significance of biases in connection with algorithmic systems?
- Have employees involved in the development and application of algorithmic systems been trained about biases?
- Are they required to reflect where biases could be present in the construction, application and evaluation of the algorithmic system?
- Does your company maintain a data and algorithm catalogue in which details on the origin of the data and the models used are stored?

Examples: In a study conducted by MIT in the field of autonomous driving („Moral Machine!“), 40 million users from more than 200 countries were confronted with the task of deciding who they would save in a dangerous traffic situation. It was noticeable that there were very few differences in the results that could be attributed to the age of the participants. However, there were clear clusters of geographical and cultural regions. Thus, it was possible to define groups of countries that could be divided into eastern, western and southern clusters. People from the southern cluster were more likely to vote for saving young people; people from the eastern cluster were more likely to save older people. For the programming of an autonomously driving vehicle, one can now ask whether one can use these results for orientation.

Some automated decisions rooted in economic interests of companies can potentially translate into biases. For example, it is common nowadays for many airlines to assign seats automatically. At first glance, this just seems to save a lot of time and effort. However, a study by the British Civil Aviation Authority (CAA) suggests that the allegedly random, automated allocation of seats, were actually deliberately dividing passengers travelling together so that they would then buy adjacent seats for a fee.[33] While this procedure is of only minor importance for the individual flight, the total number of approximately 4.1 billion people transported by airlines worldwide in 2017 alone has a much greater impact.[34]

[32] iRights.lab (2019): #algorules-Prozess, online: https://irights-lab.de/ auf-dem-weg-zu-guetekriterien-fuer-den-algorithmeneinsatz/ (accessed on: 02.02.2019).

[33] Flug-verspätet.de (2019): Airlines setzen möglicherweise Familien gezielt auseinander, online: https://www.flug-verspaetet.de/neuigkeiten/2018/11/29/airline-setzten-familien-auseinander (accessed on: 02.02.2019).

[34] Handelsblatt (2018): Weltweit mehr als 4 Milliarden Flugreisende, online: https://www.handelsblatt.com/unternehmen/handel-konsumgueter/rekord-im-luftverkehr-weltweit-mehr-als-4-milliarden-flugreisende/20862546.html?ticket=ST-334738-6qoqA6ewIhEKhkoIlgIC-ap1 (accessed on: 02.02.2019).

IV. Ethical-legal perspective on biases in algorithmic systems

As mentioned, some biases are inherent to the functioning of an algorithmic system. Thus, from an ethical and legal perspective, one of the first questions that arise is what constitutes „undesired" biases. Undesired biases can constitute unlawful, unwanted or unjustified differentiation and thus discrimination. An algorithmic system that is discriminatory according to this definition should therefore not be admissible. From an ethical-legal perspective, it is then necessary to search for solutions that counter the discrimination caused by biases in the algorithmic system. Laws represent the ethical compass of a society. In accordance with this compass, an ethical guideline must be identified which considers the social requirements in dealing with algorithms. With the help of this ethical guideline, the existing legal framework is to be examined. Where there are grey areas, additional regulatory requirements should be identified, if necessary.

Argument: There is no compelling need for new legal regulations for algorithmic systems, but rather a more effective implementation of existing regulations.

Description: To realize effective implementation, one can learn from existing regulations. According to Article 3 of the Basic Law for the Federal Republic of Germany, no person shall be favoured or disfavoured because of sex, parentage, race, language, homeland and origin, faith or religious or political opinions. No person shall be disfavoured because of disabilities. Thus, an algorithmic system that discriminates for one of the reasons mentioned above is unlawful. Systematic preferential treatment (positive discrimination or affirmative action) is still lawful and permissible if it is intentional and justified. Regarding algorithmic systems, a digital update of legal law is required. In individual cases, it would then have to be considered whether and which exceptional circumstances exist. *Handling:* The existing legal provisions of the German General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz – AGG) protect persons against discrimination. We recommend considering these against the background of technological developments and to transfer these to regulations on algorithmic systems. This means that an intensive examination of the current legal situation should take place. It is necessary to clarify how regulations are implemented and what adjustments may be necessary. In addition, the German Federal Data Protection Act (Bundesdatenschutzgesetz) already contains an initial regulation on so-called scoring[35] as well as guidelines that are intended to avoid biases by focusing on specific data. We therefore recommend that this regulation is checked for its transferability to other areas and that it is evaluated in terms of which biases are undesirable and how protection against such biases can be effectively achieved.

Example: Due to the current legal situation, it is already not permitted to exclude people who apply for a job because of their origin. It is, however, permissible to make specifications regarding language skills if this is, for example, a requirement for a profession (e.g. for the job of translating). An algorithmic system should therefore evaluate the language quality in a covering letter but should not deduce the origin of the applicant.

Argument: The handling of an algorithmic system should be dependent on the risk of discrimination and damage potential associated with it.

Description: One challenge is to reliably determine the risk of any algorithmic system in respect of discrimination and damage potential. The quantity of decisions taken should play an essential role as well as whether people are directly or indirectly affected by these decisions. Equally relevant is the dependence on this decision, for example if there is no option of switching to another provider.[36] If an algorithmic system has a low risk of discrimination and damage potent-ial, self-regulation could be a sufficient. Nonetheless, continuous monitoring of potential biases and damage should be integrated into the internal quality control. This monitoring should make both the forecast quality and the result itself the object of a detailed examination.

Handling: If an algorithmic decision-making system has a high risk of discrimination and damage potential, an external, independent evaluation should take place. In sensitive areas

[35] According to § 31 of the Federal Data Protection Act, the „use of a probability value about a certain future behaviour a natural person for the purpose of deciding on the establishment, performance or termination of a contractual relationship with this person".

[36] Katharina Zweig (2019): „Black Box Analysen zur Kontrolle von ADM-Systemen", Vortrag in der Enquete-Kommission Künstliche Intelli genz des Deutschen Bundestages, 14.01.2019, online https://www.bundestag.de/dokumente/textarchiv/2019/kw03-pa-enquete-ki/585354 (accessed on 14.02.2019).

where there is a constant risk of discrimination, constant monitoring should be introduced. For this (external) moni-toring, a procedure should be chosen which considers confidentiality interests (e.g. via in-camera procedures [37]).

These measures would allow the admissibility of certain algorithmic systems to be checked as well as simultaneously allowing their quality to be examined more closely. Thereby, the correct choice and integration of the data basis are also examined. Such monitoring measures could thus represent the minimum quality standard of algorithmic systems. Depending on their risk to discriminate and of damage potential, the given minimum standard applied could vary.

Examples: The risk of discrimination and damage potential should be used to determine by whom and on what cycle such monitoring takes place. The challenge is to reliably identify these risks. If an algorithmic system, based on previous user behavior, suggests a certain product in an online shop (e.g. a T-shirt instead of a jacket), the discrimination and damage potential is relatively low. Since it can be assumed that the company has a self-interest in making a correct forecast, self-regulation would be sufficient. Constant monitoring should be introduced in areas where discrimination against certain groups can be expected using an algorithmic system, such as in the area of human assessment and evaluation.

Argument: In cases of self-regulation and external monitoring, a minimum level is required as the benchmark for testing the specific algorithmic system.

Description: Only in this way is it possible to ensure that an algorithmic system meets the minimum standard. If the risk for damage potential and discrimination is moderate, voluntary self-commitment based on the „comply or explain“ principle could suffice.[38] In addition, the testing of confidential algorithmic systems should be carried out by a testing authority. It is important that transparency is created regarding which specific audit procedures have been carried out.

Handling: A minimum standard of algorithmic systems could be realized by certification of independent institutions. These institutions could, for example, check the database of the system, the modelling of underlying variables and the decision logic (on the bias load) of the systems.[39] An additional option is a training certificate for all those who accompany the algorithmic system in the different phases. Furthermore, an examination of the representativeness of the data – which forms the basis of learning algorithms – together with an input-output analysis could provide information about the quality of the results. In an annual report, companies could provide information on compliance with self-imposed compliance regulations for dealing with algorithmic systems. In the event of non-compliance, an external audit would be carried out.

Examples: The legal regulations on scoring, which stipulate a minimum standard (a „scientifically recognized mathematical-statistical procedure“), can be used to learn how to monitor algorithmic systems. Due to the complexity of algorithmic procedures, however, concrete minimum standards should be developed in this context. For example, it is not permissible to deduce a person‘s creditworthiness exclusively from his or her place of residence. However, the law permits a (very) high proportion of the score value to be based on the place of residence if there is scientific evidence that the use of this data leads to an accurate statement about creditworthiness.

[37] In the in-camera procedure, documents are examined by „expert panels for in-camera procedures“ established at administrative courts. The documents submitted shall not be disclosed to the public or to the parties to the dispute, nor shall they be made available to the Court of First Instance of the main proceedings. They remain in the specialised senate, i.e. „in the Chamber“. In the in-camera proceedings it shall be determined whether authorities are entitled to keep the documents secret.
Online: https://en.wikipedia.org/wiki/In_camera (accessed on: 24.02.2020).

[38] ICSA (2018): https://www.icsa.org.uk/knowledge/governance-and-compliance/features/comply-explain-uk-corporate-governance-code (accessed on: 02.02.2019).

[39] Netzpolitik.org (2018): Wie der Mensch die Kontrolle über den Algorithmus behalten kann, online: https://netzpolitik.org/2018/algorithmen-regulierung-im-kontext-aktueller-gesetzgebung/ (accessed on: 02.02.2019).

V. Outlook

To many people, the terms but also the effects associated with algorithmic systems are not yet familiar.[40] Nevertheless, ever more areas of life and work are increasingly being shaped by these systems. Demands for informed consent and digital participation can only be met if the people involved are aware of the potential effects.

Consequently, it is the responsibility of experts and decision-makers to provide information.
The biases in human decisions presented in this paper and their effects on algorithmic systems require measures in ethical, legal, socio-economic and technological areas.

This paper includes suggestions for dealing with these issues, which now need to be discussed. We recommend evaluating existing algorithmic systems against this background while, at the same time, examining the effectiveness of the proposed measures.

In addition to the focus on the topic of biases, the working group Monitoring of Algorithms has examined transparency and accountability issues as well as the question of responsibility for algorithmic systems in further papers.

[40] Initiative D21 (2020): D21-Digital-Index 2019 / 2020. The large-scale society study D21-Digital-Index provides an annual situation picture of the digital society in Germany. Online: https://initiatived21.de/publikationen/d21-digital-index-2019-2020/ (accessed on: 27.02.2020)

Initiative D21

cooperation partner:

KPMG

Working Group Monitoring of Algorithms at Initiative D21

Algorithmic systems have immense potential, particularly with regard to their growing importance in technological developments and social participation. At the same time, algorithmic systems are becoming increasingly complex and their development often lacks transparency. This creates challenges and raises various questions. In light of this, at the beginning of 2018 the Initiative D21 founded a working group to deal with issues relating to the topic of „monitoring algorithmic systems".
In the Working Group Monitoring of Algorithms at Initiative D21 relevant issues were discussed by interdisciplinary experts from three perspectives: technological, socio-economic and ethical-legal. The technological perspective refers to the practical feasibility of Monitoring of Algorithms and deals with the conditions, problems and possibilities. The socio-economic perspective determines the social and economic opportunities and challenges posed by the application of algorithmic systems and how risks can be counteracted. The ethical and legal perspective deals with the development of a legal base to ensure the fair regulation of algorithmic systems.

Theses were derived from the discussions and published in three Essays on Digital Ethics: „Bias in algorithmic systems", „Transparency and Explainability of algorithmic systems" and „Responsibility for algorithmic systems". As a summary, 9 guidelines for monitoring algorithmic systems have been developed. These recommendations contain suggestions as to which regulations of algorithmic systems might be ethically necessary, how these affect society and the economy, and how they could be implemented technologically. They include basic questions for further discussion and serve as a call to action for continuous review and further development in this area.

MONITORING
OF ALGORITHMS

IMPRINT

Initiative D21 e.V.
Reinhardtstraße 38
10117 Berlin
Germany
www.InitiativeD21.de

Phone: 0049 30 5268722-50
kontakt@initiatived21.de

Download
initiatived21.de/publikationen/denkimpulse-zur-digitalen-ethik