Initiative D21

cooperation partner: KPMG

Working Group Monitoring of Algorithms | Translation: 27 January 2020
Original German version published on 18 June 2019

DENKIMPULS DIGITALE ETHIK:

Transparency and explainability of algorithmic systems

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

COLLABORATORS Aljoscha Burchardt (German Research Center for Artificial Intelligence - DFKI), Lena-Sophie Müller (Initiative D21 e.V.), Dr. Nora Schultz (Office of German Ethics Council), Prof. Barbara Schwarze (Initiative D21 e.V./ Compe-tence Center Technology-Diversity-Chancengleichheit e. V.), Staff at KPMG AG and KPMG Law

- **Algorithmic systems play an increasingly significant role in all areas of life. The basis for the testability of these systems is transparency and explainability.**
- **Transparency describes the provision of information and does not necessarily lead to common understanding. On the contrary, too much information can limit comprehension. Nevertheless, transparency is a necessary basis for explainability.**
- **Information only becomes explainable when the connections that led to its creation and its immediate effects can be understood by a person. Consequently, explainability requires the target group-specific provision of information.**
- **It is imperative that algorithmic systems are designed in such a way that they can be comprehended by a given actor. Depending on their context, requirements for transparency and explainability can vary.**

I. Introduction

Transparency and explainability: definitions and explanations

Few other developments have changed our lives in recent years as much as digitalisation, i.e. the introduction of information technology systems in all areas of life. Computer programs are based on innumerable algorithms, which often perform a wide range of tasks at different levels. There are varying complexities of algorithms: for the adding and subtracting of numbers, for the creation of a file on a hard disk, for the sorting of lists or the training of a machine learning system. The complexity of these algorithmic systems can be seen, among other things, in the length of the necessary program code, which can range from a few lines to entire program libraries. In the following paper, the term „algorithmic systems“ refers not only to the program code, but also to the various stages of an algorithmic system, including processes of data selection and evaluation, decisions on interface design, and access options for people interacting with the system.

Algorithmic systems have long been used in countless areas and facilitate a multitude of everyday tasks. They

accompany people through life – often unnoticed – from admission to a certain school or university, through job application processes, possible promotion to the assessment of creditworthiness.[1] Nevertheless, there is often ignorance among people who consciously or unconsciously use algorithmic systems. On the one hand, there is distrust of algorithmic systems[2], for example, in respect of which basic decisions are made by systems that produce recommendations. On the other hand, there is often blind trust in technology.[3]

Algorithmic systems are not independent entities that act on their own initiative.[4] The more they are perceived as a pivotal element in decision-making processes, and the less explainable they become due to their complexity, the greater the mistrust and the desire for monitoring. Very few people understand the criteria underlying the calculations performed by these systems. In order to strengthen confidence in algorithmic systems, it is therefore necessary to make them more transparent and thus more explainable. Algorithmic systems currently offer us an opportunity to make decisions that were previously subject to human discretion, more transparent, fairer and less biased. This has great potential.

This paper examines how transparency and explainability relate to one another in different applications of algorithmic systems and how they differ for different groups of people. Relevant questions were identified and dealt with from a socio-economic, technological and ethical-legal perspective. The paper was prepared within the framework of the working group Algorithm Monitoring of the Initiative D21 e.V. The arguments presented for the handling of transparency and explainability contribute to a more differentiated debate and are intended to initiate a broader discussion on the meaning and purpose of transparency and explainability in algorithms. The aim is to create a further basis for sustainable use of the terms as well as measures to create transparency and explainability.

The term **transparency** merely describes the provision of information. The concept of **explainability**, on the other hand, contains a subjective element and depends on whether information is presented in appropriate language and takes the background knowledge and intellectual abilities of addressees into account.[5] In this respect, transparency is a prerequisite for explainability. Information becomes explainable when the connections that led to the creation of information and its immediate effects are understood by a given person. Transparency and explainability can be in conflict with each other: The EU General Data Protection Regulation (GDPR) requires companies to provide information about an individual on request. This enables people to request and view the data stored about them. However, they can expect to receive an extensive list of individual data points, bereft of any elaboration or explanation.[6] This may be understood as transparency, but it does not guarantee explainability. An overload of information can obstruct explainability. Comprehension requires contextualisation, the presentation of connections and an explanation of possible effects in suitable language and scope. This requires target-group-specific preparation of the corresponding information. Depending on the expertise of the addressee, a different type of information provision is required.

1 O'Neil, C. (2016): Weapons of Math Destruction; Crown Random House; Online: https://weaponsofmathdestructionbook.com (Accessed: 14.06.2019)
2 Vodafone (2016): Big Data; Online: https://www.vodafone-institut.de/wp-content/uploads/2016/01/VodafoneInstitute-Survey-BigData-en.pdf/ (Accessed: 12.06.2019)
3 Fink, R.D.: Vertrauen in autonome Technik; Online: https://core.ac.uk/download/pdf/46915729.pdf (Accessed: 12.06.2019)
4 OECD Principles on Artificial Intelligence; Online: https://www.oecd.org/ going-digital/ ai/ principles/ (Accessed: 11.06.2019)
5 Schmitt, A. (2005): Bedingungen gerechten Handelns. Motivations- und handlungstheoretische Grundlagen liberaler Theorien; Springer VS, S. 105
6 Nocun, Katharina (2018): Die Daten, die ich rief. Verlag: Bastei Lübbe; weiterführend: http://kattascha.de/worum-geht-es-im-buch-die-daten-die-ich-rief/ (Accessed: 14.06.2019)

Actors within and outside the algorithmic system

In the following, we will introduce the different actors that need to be considered within and outside algorithmic systems. These are summarised in the following illustration in four levels with different objectives with regard to transparency and explainability. People who knowingly or unknowingly use the algorithmic system are referred to as **users**. People that develop, test and/ or implement algorithmic systems are called **designers**. In addition, there are people who make decisions on the commissioning and deployment of algorithmic systems (**decision-makers**), as well as experts from various disciplines acting as (external) **auditors** of algorithmic systems.

Users

Persons may be affected as active users of an algorithmic system or as indirect objects of data generation. Active use, for example, would then concern users of a route-planning software. Indirectly affected persons are, for example, persons whose movement data from mobile devices is collected and used for the training of said route planning software, which they themselves may not be using.

For **users**, explainability is measured by the extent to which a given target group can understand the implications of a given algorithmic system. Measures to ensure explainability are generally planned for the majority of people in the respective target group. Special attention should also be paid to the participation of people with special requirements for explanations, such as children and people with disabilities.[7] Custom-made systems for specific target groups, e.g. specialist software for medical personnel, need initially only be explainable to them. For example, an explanation in medical jargon is helpful for medical personnel. In order to further understand information about the use of algorithmic systems, it is helpful to offer generally accessible further education courses[8] and to expand school education to include more basic information courses on statstics and computational thinking.[9]

A first step towards explainability is a transparent account of the use of algorithmic systems supplemented by a short explanation. This could, for example, be analogous to the common wording „This letter was generated automatically and is valid without a signature“. For example, chats should clearly indicate which questions are answered by chatbots and how a human can be reached. In certain situations, information needs to be collected quickly, for example during an (online) purchase. A simplification of the representation and a high degree of comparability of this information can be made possible by structuring.[10]

After the use of an algorithmic system, the criteria for the evaluation, their weighting and the data on which the calculation of the result is based might be disclosed to those affected. For example, when an algorithmic system calculates a person‘s creditworthiness, that person should be able to understand which data affected the outcome and how.

In addition, as it cannot generally be assumed that users possess the knowledge and expertise necessary to investigate suspicions of discrimination against data subjects, an external body could be tasked with pursuing this on their behalf. To ensure that appeals are accessible to all people there must be a concrete external means of appeal in addition to an internal complaint management system.

7 Bundeszentrale für politische Bildung (2018): Allgemeine Geschäfts-Bedingungen der bpb in leicht verständlicher Sprache; Online: https://www.bpb.de/shop/201038/allgemeine-geschaefts-bedingungen in leichter Sprache/ (Accessed: 08.05.2019)

8 Elements of AI: Free online course; Online: https://www.elementsofai.com/ (Accessed: 08.05.2019)

9 GI (2019): GI begrüßt Forderung nach Informatik-Pflichtfach in Niedersachsen; Online: https://gi.de/meldung/gi-begruesst-forderung-nach-informatik-pflichtfach-in-niedersachsen/ (Accessed: 08.05.2019)

10 Transparency International Deutschland e.V.: Initiative Transparente Zivilgesellschaft; Online: https://www.transparency.de/mitmachen/initiative-transparente-zivilgesellschaft/ (Accessed: 08.05.2019)

Designers

Designers are people who conceptualise, develop, test and/or distribute an algorithmic system. Companies in this field will weigh the costs against the benefits of fulfilling the criteria of transparency and explainability. Costs for more complex procedures must be justified. Internal developments are regarded as trade secrets worthy of protection. Data, algorithms and other factors should be readily available for internal testing and quality improvements. Thus, an internal quality control can insist on internal transparency to improve processes and procedures throughout the various phases of an algorithmic system's lifecycle. Such traceability within the system serves to monitor functionality and can also be used for improvement.

Decision-makers

Decision-makers are legally, technically or politically responsible persons who decide which algorithmic systems are to be used for which purpose and how they are to be tested. In order to make balanced decisions and to develop guidelines, regulations and laws for algorithmic systems, an interdisciplinary exchange, as well as the integration of different cultural backgrounds is imperative.[11] Areas affected by regulations include the planning of data-saving or data-intensive algorithmic systems. Such decisions must be made at an early stage, as a later change is much more difficult and expensive. Analogous to „Privacy by Design", requirements for „Transparency by Design„[12] or "Explainability by design"[13] can be formulated.

The requirements at the level of the decision-makers concern the explainability of the processes, results and limitations of algorithmic systems. An obligation to carry out external audits would lead to greater explainability. A „balance between openness and secrecy" remains important for society.[14] The rights of the various actors must therefore be identified and, if necessary, protected. For example, those affected may have a right to secrecy in certain situations.[15]

External auditors

From the point of view of external auditors, many factors must be taken into account. External means here that the persons are not part of the group of persons designing the algorithmic system. Instead, they review the system as external third parties. Relevant data, algorithms, models and processes should be disclosed for a comprehensive examination, so that possible mistakes and biases/discrimination resulting from the algorithmic system can be discovered. Explainability of algorithmic systems can only be achieved through interdisciplinary work: for example, technical experts look at the source code, legal experts assess the explainable behaviour, and communication experts subsequently explain the given results in a way that is easily understood by the target audience. The results of external audits should be published.[16] External auditors also include those who review the extent to which laws and regulations are being implemented.

11 Balkow, C., Eckardt, I. (2019): Bias in algorithmic Systems: Online: https://initiatived21.de/publikationen/denkimpulse-zur-digitalen-ethik/
12 Mascharka, et.al (2018): Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning; Online: http://openaccess.thecvf.com/content_cvpr_2018/papers/Mascharka_Transparency_by_Design_CVPR_2018_paper.pdf (Accessed: 23.05.2019)
13 Stackpole, Beth (2020): 5 steps to 'people-centered' artificial intelligence; online: https://mitsloan.mit.edu/ideas-made-to-matter/5-steps-to-people-centered-artificial-intelligence (Accessed: 13.01.2020)
14 Denk, Felix (2016): Wann Algorithmen transparent sein sollten – und wann nicht; Online: https://www.fluter.de/kuenstliche-intelligenz-ethik-algorithmen-kontrollieren/ (Accessed: 08.05.2019)
15 Duttge, G., et.al. (2015): Normatives Fundament und anwendungs-praktische Geltungskraft des sogenannten Rechts auf Nichtwissen; Online: http://www.recht-auf-nichtwissen.uni-goettingen.de/ (Accessed: 13.06.2019)
16 Semsrott, A. (2019): Verwaltungsgerichte entscheiden für Transparenz; Online: https://fragdenstaat.de/blog/2019/05/21/topf-secret-klagewelle/ (Accessed: 13.06.2019)

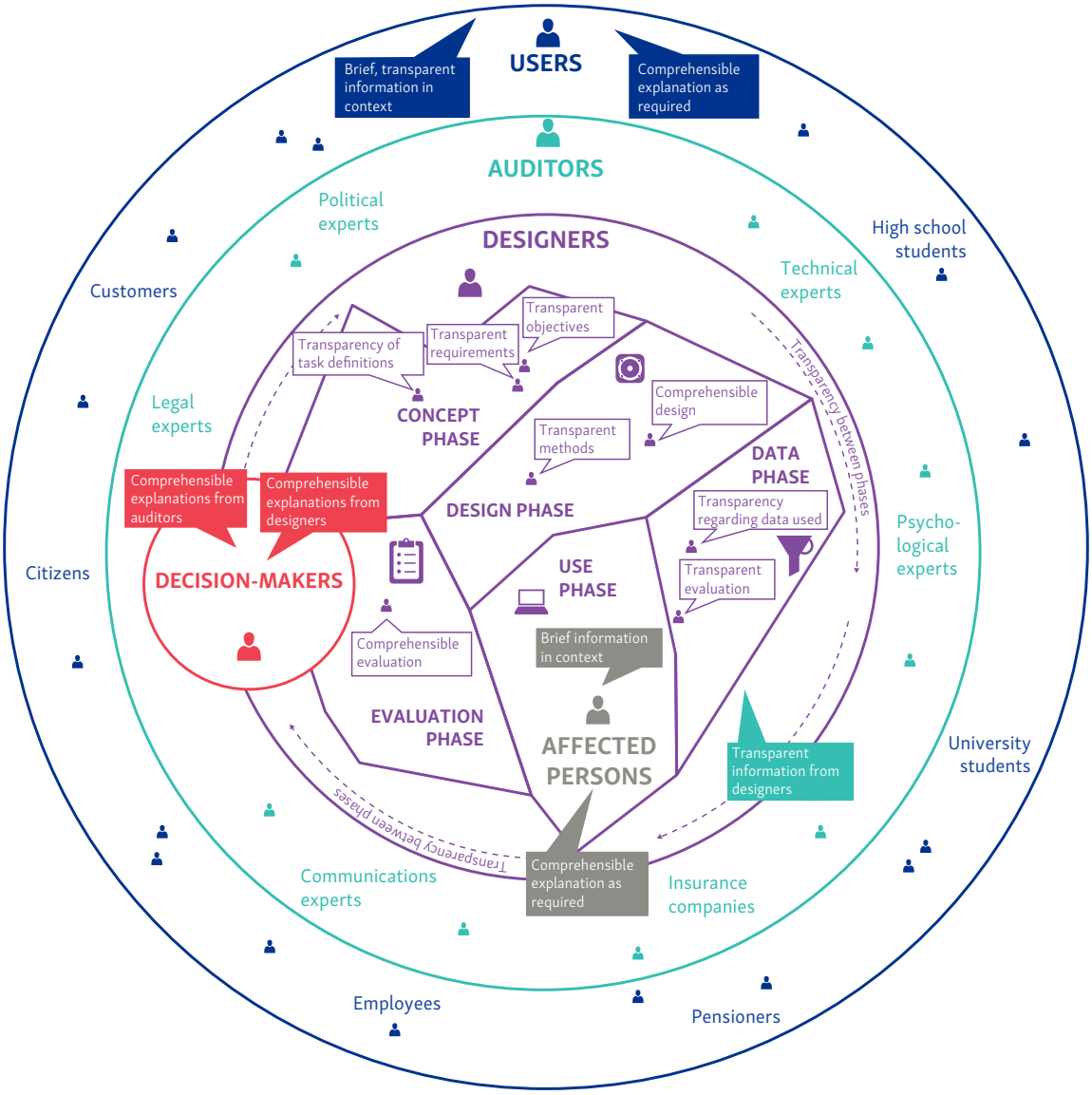

Figure 1: Relationship between transparency and explainability regarding the four groups mentioned within and outside a given algorithmic system.

Designing algorithmic systems requires transparency within the system

The inner circle of the graphic shows the development cycle of an algorithmic system. In the concept phase, requirements for outputs and objectives are defined. These requirements are implemented in the design phase. In the data phase, data is collected, evaluated and prepared for use. In the deployment phase, users come into direct contact with the system through their use of the system. In the evaluation phase, the results are evaluated. Thus, the **designers** of an algorithmic system are involved in all phases of the system. A transparent exchange of information can take place between the phases. However, comprehensible explanations are often sufficient for agreements.

Decision-makers must be able to understand the effects and consequences of algorithmic systems.

In order to assess the various implications of algorithmic systems, decision-makers must communicate with each other as well as with the other actors. Depending on their own expertise, they need transparent information

or comprehensibly processed information by experts from other disciplines or specialisations. Depending on the context of use, a decision must be made as to whether self-learning algorithms, whose outputs are more difficult to comprehend, are appropriate. When assessing algorithmic systems, not only the algorithms themselves have to be considered, but also the data sources used, the implementation process and future types of usage should be assessed. Furthermore, it must be communicated to what extent the algorithmic system is used in supporting decision-making versus as the sole basis for decisions.

External auditors require insight into the entire development cycle of the algorithmic system

External auditors sometimes stand between designers, decision-makers and those affected. For a meaningful audit, they need transparency over the entire life cycle of an algorithmic system, the decision-making processes and the needs of those affected. They provide an explainable assessment to the audit client – a decision-maker or an affected person. The algorithmic system is thereby audited, for example, for possible discrimination. By publishing audit reports, explainability can be achieved without complete transparency of the algorithmic system.

Users need comprehensibly processed and explained information

Those directly affected are active users of the algorithmic system. So that those directly affected know that they are currently using and engaging with an algorithmic system, they should be notified by the system in the context of using the application. This could be made possible using symbols. If necessary - for example to understand whether they have been wrongfully treated by algorithmic systems - an explainable explanation of the decision-making process must be readily available.

Due to the general dissemination of algorithmic systems, persons may be affected by algorithmic systems indirectly without actively engaging with them e.g. automated border controls.[17] Indirectly affected users are unknowingly part of the system because their analogue data have been digitised or their digital footprints are used. It is important to consider how indirectly affected users can be informed when algo-rithmic systems are used. For example, it could be shown when and how it is appropriate to use the personal data of indirectly affected users to improve an algorithmic system.

Different requirements for transparency and explainability

In the following section, relevant questions concerning the different actors are examined from three perspectives: technological, socio-economic and ethical-legal. The technological perspective refers to the practical feasibility of requirements of transparency and explainability. It deals with the conditions, problems and possibilities associated with providing transparency and explainability. The socio-economic perspective determines which social and economic opportunities and challenges arise from what is required of algorithmic systems in terms of their transparency and explainability and how these challenges can be met. The ethical-legal perspective deals with the development of legal foundations and a possible regulation of algorithmic systems on the basis of their risk for discrimination and damage.

17 Automated border control system EasyPASS; Online: https://www.easypass.de (Accessed: 11.06.2019)

II. Technological perspective on transparency and traceability in algorithmic systems

The demands for a better understanding of algorithmic systems are constantly increasing due to their widespread use. Generally prevailing distrust of extensive data collection[18], unfathomable results of algorithmic systems[19] and decisions supported by artificial intelligence[20] often lead to the demand for mandatory transparency. As explained above, transparency does not automatically lead to comprehension. Greater transparency of algorithmic systems can, however, facilitate access to relevant information, in principle enabling explainability of the results and permitting a better evaluation of the systems. Target-group-specific processing of information would de facto lead to more explainability and thus permit better evaluation of the systems. Transparency is therefore necessary for monitoring and explainability serves as a prerequisite for comprehension.

Argument: Transparency can serve to improve algorithmic systems

Description: From a technological perspective, transparency is - with certain limitations in machine learning - technically viable. However, it is questionable whether this transparency can be converted into explainability. As an intermediate stage, explainability means that an expert can recognise what is happening in an algorithmic system. „Explainable AI" (abbreviated XAI[21]) describes the field of research aimed at making the decisions of an artificial intelligence - an algorithmic system – easily explainable, i.e. to clarify why and how certain factors were weighted in algorithmic systems and which assumptions form the basis of a given decision. The algorithmic system therefore must be planned in such a way that outputs are made readily available for analysis.

Example: Research on explainable AI emphasises the usefulness of assessing how successfully an algorithmic system has learned. For example, researchers at TU Berlin demonstrated that AI systems often learn from features that are not visible to humans. For example, boats and trains were not recognised by their structure, but by the background in the photos the algorithmic system learned from: boats were recognised by water; trains were recognised by the fact that there were rails. Images of horses came from a specific database and were recognised by the algorithmic system on the basis of a copyright feature.[22] The issue of possible discrimination raises the technical question of whether the systems identify certain features that serve as substitutes for prohibited personal features, e.g. when parenthood is inferred from gaps in women's CVs.

Handling: Previously used algorithmic systems have hardly produced any output on their calculation methods. Instead, they are used in the background.[23] However, the GDPR requires explanatory information.[24] Complaint management can provide indications of where improvements might make sense. In encryption techniques, the disclosure of the source code is regarded as a quality feature.[25]

Argument: Publishing too much data can inhibit explainability

Description: Due to the ever-increasing use of technological solutions as well as ever-increasing storage capacity,

18 Joler, V., Petrovski, A. (2016): Immaterial Labour and Data harvesting; Online: https://labs.rs/en/facebook-algorithmic-factory-immaterial-labour-and-data-harvesting/ (Accessed 12.06.2019)
19 Prof. Dr. Zweig, K. et. al (2018): Wo Maschinen irren können; Online https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WoMaschinenIrrenKoennen.pdf (Accessed: 12.06.2019)
20 Canon, G. (2019): How Taylor Swift showed us the scary future of facial recognition; Online: https://www.theguardian.com/technology/2019/feb/15/how-taylor-swift-showed-us-the-scary-future-of-facial-recognition/ (Accessed: 08.05.2019)
21 Eisenstadt, V., Althoffl, K.-D., (2018): A Preliminary Survey of Explanation Facilities of AI-Based Design Support Approaches and Tools, Online: https://www.dfki.de/fileadmin/user_upload/import/9983_LWDA_2018_paper_59.pdf (Accessed: 12.06.2019)
22 Lapuschkin, Sebastian et al. (2019): Unmasking Clever Hans predictors and assessing what machines really learn; Online: https://www.nature.com/articles/s41467-019-08987-4.pdf/ (Accessed: 08.05.2019)
23 West, S.M., Whittaker, M. and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Online: https://ainowinstitute.org/discriminatingsystems.html (Accessed: 14.06.2019)
24 Selbst, Andrew D; Powles, Julia (2017): Meaningful information and the right to explanation; Online: https://doi.org/10.1093/ idpl/ ipx022/ (Accessed: 08.05.2019)
25 Pillitteri, V.; Lightman, S. (2015) Guide to Industrial Control Systems (ICS) Security; Online: https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-82r2.pdf (Accessed: 14.06.2019)

more data is available today than ever before. Algorithmic systems can now handle increasingly complex tasks, but they also require immense amounts of relevant data to do so. The processing of such data, which needs to be sorted, ordered and sometimes supplemented, accounts for approx. 80% of the effort involved in training neural networks. This important task is often outsourced[26], which may pose additional challenges for an explainable analysis. Accuracy and comprehensibility are in tension with each other when it comes to achieving required explainability. The more technical details are disclosed, perhaps down to the basics of the hardware or software, the more difficult it becomes for most of those affected to understand them.

Example: Anyone who makes a purchase in an online shop must provide various personal information: name, address, date of birth, account details. The data protection regulations describe in detail what is used for what purpose. When the provider's data is queried, some of the requesters receive detailed spreadsheet documents including system data, device types and operating systems used.[27]

Handling: Depending on the target group and context, a gradation into different levels of transparency should take place:

Users: People who use the algorithmic system in everyday life usually require less depth of detail. An important piece of information that can also be captured quickly would be a clear indication that an algorithmic system has been used, combined with a brief explanation. This can be supplemented with core information about its nature or function.[28] Simple explanations of a given displayed result facilitate understanding.[29] If desired, a detailed explanation must be easy to find. Finally, it should be possible to object to erroneous results.[30] Possible solutions therefore include options that offer as much diversity of information as possible which would address all possible questions to an interested user, as well as abstractions that allow a brief overview similar to food labelling.[31]

Decision-makers: People who commission or use algorithmic systems need explainable explanations for all phases of the algorithmic system. Not entirely transparent algorithmic systems can also be evaluated through certification of processes in planning, development and use.[32]

Auditors: People who develop, test or audit algorithmic systems need as much information as possible in order to assess how reliably the system achieves the desired results.

26 Stouffer, Keith et. al (2015): The invisible workers of the AI era - A new type of blue-collar industry has emerged around curating the data that powers AI; Online: https://towardsdatascience.com/the-invisible-workers-of-the-ai-era-c83735481ba/ (Accessed: 08.05.2019)
27 Heidrich, J; Maekeler, N. (2019): Antwortet uns! DSGVO-Datenauskunft im Selbsttest; c´t 2019, Heft 13, S.171
28 Bier, C. (2018) Umsetzung des datenschutzrechtlichen Auskunftsanspruchs auf Grundlage von Usage-Control und Data-Provenance-Technologien; Online: http://publica.fraunhofer.de/dokumente/N-484529.html (Accessed: 14.06.2019)
29 Ortloff, A.M. et al (2018): Evaluation kontextueller Datenschutzerklärungen; Online: http://publica.fraunhofer.de/dokumente/N-528032.html (Accessed: 14.06.2019)
30 McLeod, A. (2017): How I turned a traffic ticket into the constitutional trial of the century; Online: https://arstechnica.com/tech-policy/2017/01/op-ed-how-i-turned-a-traffic-ticket-into-the-constitutional-trial-of-the-century/ (Accessed: 24.02.2020)
31 Creative Commons Licenses; Online: https://creativecommons.org/use-remix/cc-licenses/ (Accessed: 24.02.2020)
32 KI Verband (2019): KI-Gütesiegel – AI Made in Germany; Online: https://ki-verband.de/ki-guetesiegel-ai-made-in-germany/ (Accessed: 08.05.2019)

III. Socio-economic perspective on transparency and explainability in algorithmic systems

From a socio-economic point of view, assessing transparency and explainability requires looking at it in various contexts and from the perspective of various actors within the socio-economic sphere. Such actors could belong to the following groups: government organisations, business enterprises, the intermediary sector (such as churches, research institutions and associations) and citizens.

The state, through public administration, pursues public agendas, i.e. tasks that are explicitly subject to the common good. The state can legally require a given person to act on or refrain from certain behaviour, for example by means of an official building permit or a penalty ticket. Since the government can intervene extensively in the rights of third parties, decisions must be transparent to the public in order to be verifiable. Commercial enterprises, on the other hand, are not obliged to be transparent in their actions in all areas for reasons of competitive advantage (trade secrets). This should serve the creation of value and the preservation of material prosperity. Within the framework of the rule of law, companies can be obliged to prove that, for example, individual persons/ groups of persons are not harmed, disadvantaged or exploited.

Business models based on data-generation as well as digital products and services are indispensable today. Within this framework, the individualisation of products is increasingly becoming the focus of attention. Whether personalised film suggestions, language assistants or genetic tests, they all have the goal of being personally tailored to different people. This development has led to the ever greater generation of data, as well as subsequent processing of such data by algorithmic systems. The explainability of such algorithmic systems therefore also requires an understanding of how personal data is used. It is also important to check whether the algorithmic systems used makes correct as well as fair decisions. The next section examines the views of citizens and the perspectives of state organisations and business enterprises.

Argument: Complete transparency of algorithmic systems is not necessarily useful

Description: Generally requiring actors to make their data, models and algorithms publicly accessible and completely transparent is not the answer to the question raised here about the correct handling of explainability. It is important to weigh up how and for whom transparency or explainability should be created. There are risks that comprehensive transparency could be exploited for personal or criminal purposes and could result in economic or social damage.

Example: When using search engines, knowledge of the exact algorithm could lead to certain products being placed more prominently than is relevant for the searcher. Distorted search results and the disadvantaging of smaller websites lead to economic damage.

Handling: Especially for the example above, search engine designers could be required to name influencing factors for the algorithm (e.g. optimal formatting or unnatural link patterns), but not to disclose the exact model or decision logic behind it (e.g. the exact weighting of optimal formatting or the recognition of unnatural link patterns in indexing). In the first step, a rough segmentation or classification of actors (e.g. government bodies, financial institutions, etc.), considering their core tasks or business models, should be focused on clarifying algorithm-based decisions. More segmentation could result as regards the extent to which a given algorithmic system portrays a particularly high damage potential with the disclosure of its algorithms. In the next step, it would be possible for those who are committed to greater transparency to carry out a context-specific risk assessment analysis. In a third step, it must be determined what is to be regarded as the minimum measure of explainability.[33]

Argument: Explainability is more than technical transprency

Description: It is not enough to make information or data transparent in order to be able to understand this

33 Prof. Dr. Zweig, K. (2019): Algorithmische Entscheidungen: Transparenz und Kontrolle; Online: https://www.kas.de/ c/ document_library/

information. It must be explained and put into context. Explainability in algorithmic systems thus results from an interdisciplinary approach. In addition to the underlying data and algorithmic models, non-technical aspects, such as the context of application, the economic motivation behind it or the underlying ethical assumptions, including biases, must also be considered.

Example: When granting loans or determining the credit-worthiness of a given individual, different data sources are used. Among other data points, personal socio-economic data such as place of residence and social status are used and can thus lead to different results. Explainability can be established if there is a justification as to why these data points are used and whether or how they are included in the evaluation. Using the risk assessment analysis mentioned above, the influencing factors and the motivation behind them could be identified.

Handling: Context-specific risk assessment analyses could be introduced to establish a minimum degree of explain-ability, as is already customary in the health sector.[34] Technical, political, economic and ethical aspects should be explicitly addressed here. Examples would include:

- "Are socio-economic characteristics of individual persons, such as age or mother tongue, included in the data and do they have a direct impact on the results of the calculations?"
- „For what reason were these characteristics chosen? How were the queries integrated?"
- "What options are offered to those affected to change or even delete characteristics?".

The answers could be made available to the public without the models or algorithms themselves necessarily being published. The associated standardisation, explanation to a suitable extent and comprehensible language would make it easier to understand them.

Argument: Standards serve to make algorithmic systems comparable

Description: Designers of algorithmic systems need standards so that they can comply with obligations of explainability. Standards defined by decision-makers can also serve those affected or auditors who, for example, are reviewing the explainability of an algorithmic system. Based on these criteria, it can be evaluated better and faster whether it is a comprehensible system or not.

Example: Companies that use algorithmic systems are obliged to publish corresponding general terms and conditions and to draw attention to data protection and data use. Due to the scope and complexity of these general terms and conditions, it is often impossible for users to understand how and for what purpose personal data is used. Through certification of processes in the planning, development and deployment periods, even algorithmic systems that are not completely transparent may be evaluated by users through the corresponding labels.[35]

Handling: A possible standard could, for example, be docu-mentation in a language that is suitable for the persons con-
cerned, so that it is immediately apparent which data are used by which user groups in the algorithmic system. Further-more, quality manuals for statistics such as the „Qualität-shandbuch der Statistischen Ämter des Bundes und der Länder"[36] could be used for the evaluation of analyses.

get_file?uuid=533ef913-e567-987d-54c3-1906395cdb81&groupId=252038 (Accessed 14.06.2019)

34 Bundesinstitut für Risikobewertung (2010) Leitfaden für gesundheitliche Bewertungen; Online: https://mobil.bfr.bund.de/ cm/ 350/ leitfaden-fuer-gesundheitliche-bewertungen.pdf (Accessed 14.06.2019)

35 KI Verband (2019): KI-Gütesiegel – AI Made in Germany; Online: https://ki-verband.de/ ki-guetesiegel-ai-made-in-germany/ (Accessed: 08.05.19)

36 Destatis (2018): Qualitätshandbuch; Online: www.destatis.de/ DE/ Methoden/ Qualitaet/ Qualitaetshandbuch.pdf?__blob=publicationFile/ (Accessed: 03.04.19)

IV. Ethical-legal perspective on transparency and explainability in algorithmic systems

From an ethical-legal point of view, **transparency** is understood as the general provision of information without any claim to the fact that this information has been processed. This information is made explainable and comprehensible through target-group-specific preparation. This does not only mean technical explainability, but also ethical, psychological or economic comprehensibility. This means that users and decision-makers should not only understand the technical functionality of an algorithmic system, but should, for instance, also be able to evaluate its effects. To this end, it must be possible to question, understand and examine relationships.

The provision of information must be adapted to the target group to which it is directed. In general, a distinction must also be made here between users, designers, decision-makers, and external auditors. This is because, while external auditors can derive comprehensibility with a lot of information, common users often lack such technical competence and the same information must therefore be prepared in a more understandable way. Mere disclosure of all information (e.g. to meet a legal requirement) would have the opposite effect on this group of users; they would be overwhelmed by the amount of information. An excess of information could even lead to essential information being neglected.

From an ethical-legal point of view, **explainability** enables people to assess and evaluate the circumstances. In terms of transparency, the GDPR[37] also makes a reference to explainability. For example, according to the GDPR, information is only transparent if it is precise, easily accessible and easy to understand. If children are affected, the information must be provided in a language suitable for children. The greater the number of actors and the greater the technological complexity, the more precise and easier to understand the available information must be. However, many applications of algorithmic systems are not covered in the GDPR.

Those affected must be informed if a fully automated decision-making process or a classification of individuals (profiling) takes place and what consequences this has. This, however, leads to many grey areas, as the demarcation to merely preparatory algorithmic systems is extremely inaccurate and dependent on individual cases. If, for example, an algorithmic system is used to sort applications, it is questionable whether people see this merely as preparation for a decision or as the decision itself. Although the system does not decide automatically, it is used to speed up and simplify the decision-making process. In addition, the GDPR does not contain any regulations on the effects of algorithmic systems if no persons, but only machines or commercial traffic are affected.

Against this background of ethical-legal considerations, there is a need for action regarding the transparency and explainability of algorithmic systems.

37 Regulation (EU) 2016/ 679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with re gard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation) Online: https://eur-lex.europa.eu/ legal-content/ EN/ TXT/ ?qid=1580129774333&uri=CELEX:32018R1725 (Accessed: 27.01.20)

Argument: Algorithmic systems should be evaluated in their specific context and based on their risk of discriminating and causing damage

Description: The evaluation of algorithmic systems cannot be carried out based on general risk classes, but must always consider the specific context of use, as this has a decisive influence on the respective risk of discrimination and damage.

Example: An algorithmic system used for profiling terrorists has a high risk of discriminating and causing damage.[38] The need for transparent and explainable processing would therefore be greater than for an algorithmic system used to control a production line.

Handling: For effective implementation, an evaluation of the field of application could be determined based on certain criteria. When determining the risk for discrimination, it should be taken into account whether people are directly or indirectly affected by the result of an algorithmic system and whether the system categorises individual people. Equally relevant is the dependency on this decision, for example, if it is not possible to switch to another provider.[39] The risk could be composed of possible economic, psychological and ecological damage on the one hand, and the number of people affected on the other. Governmental guidelines, which determine the risk of discrimination and damage, could serve as a guide.

Argument: A general obligation to make algorithmic systems fully explainable at any time and in any context leads to over-regulation

Description: The effort that must be put into creating complete explainability at any time, in any context and for any target group is immense.[40] Accordingly, it must be assessed whether, for whom and to what extent explainability must be guaranteed.

Example: In an algorithmic system used to control a production line, the risk of discrimination and damage is low, since initially no people are affected. In the event that the wrong decision is made, the damage would primarily be borne by the company. Accordingly, the effort that would have to be made to make this algorithmic system explainable at any time and in any context would be disproportionate. However, the given company would be free to contractually demand a higher degree of transparency or explainability from the provider or developer.

Handling: Efforts for and benefits of extensive transparency should be considered. To what extent an explanation should take place depends on the risk of discrimination and damage. In addition, it should be defined for whom the system must be transparent and comprehensible. For example, algorithmic systems with a high risk of discrimination and damage would always have to make a comprehensible explanation easily available for both users as well as decision-makers. Algorithmic systems with a low discrimination and damage risk could be comprehensible for decision-makers, designers and auditors. In the case of safety-critical systems, the system should only provide comprehensible processing for decision-makers and auditors of state-approved institutions.

Argument: Algorithmic systems should always be transparent, although not for everyone

Description: For explainability to be achieved in the event of a mistake or an external evaluation, transparency must be ensured during the entire development cycle of an algorithmic system. For whom and to what extent this transparency should then lead to comprehensibility depends on the risk of discrimination and damage.

Example: A fitness tracker, which records an individual's movement pattern, would appear to have a comparably low discrimination and damage risk. The evaluation of this data could, however, lead to health-endangering behaviour or later to refusal from health insurance providers. Before this happens explainability must be guaranteed.

Handling: Documenting the goals, data, methods, test and release processes assures not only higher quality, but also guarantees more transparency. Depending on the discrimination and damage risk of an algorithmic system, a

38 Bilger, A. (2018): Künstliche Intelligenz in der Verbrechensbekämpfung; Online: https://www.boell.de/de/2018/01/29/kuenstliche-intelligenz-der-verbrechensbekaempfung/ (Accessed 13.06.2019)

39 Prof. Dr. Zweig, K. (2019): Algorithmische Entscheidungen: Transparenz und Kontrolle; Online: https://www.kas.de/c/document_library/get_file?uuid=533ef913-e567-987d-54c3-1906395cdb81&groupId=252038 (Accessed 14.06.2019)

40 DFKI (2017): Künstliche Intelligenz; Online: https://www.dfki.de/fileadmin/user_upload/import/9744_171012-KI-Gipfelpapier-online.pdf/ (Accessed: 08.05.2019)

decision would be made for whom and the extent to which this transparency applies as well as to whom it must be explainable. Such data transparency would not mean that operators would have to publish all information at any given time, but would have to provide relevant information on request, for example in the event of a complaint.

Self-documentation mechanisms and logs could serve as a minimum standard. It is important that in the event of an audit enough data was collected to carry out a post-hoc analysis. Operators of an algorithmic system with a high risk of discrimination and damage would have to ensure not only internal transparency, but also external traceability. The assessment of the discrimination and damage risk would have to be continuously reviewed and, if necessary, changed. The further development of algorithmic systems creates possibilities that were not foreseeable at the time of the initial development or first use.

V. Outlook

Many people are not yet familiar with the concepts and effects of algorithmic systems.[41] Demands for informed consent and digital participation can only be met if users and decision-makers are aware of the effects. It is the responsibility of experts to act in an informative manner.

The arguments presented in this paper require measures in ethical-legal, socio-economic and technological areas. Suggestions for how to tackle these issues are mentioned, which now need to be discussed further. We recommend evaluating existing algorithmic systems against this background and at the same time examining the effectiveness of the proposed measures.

In addition to the focus on the topic „Transparency and Explainability of algorithmic systems“, the working group Algorithmic Monitoring has already addressed the topics „Bias in algorithmic systems“ and „Responsibility for algorithmic systems“.

41 D21-Digital-Index (2018/ 2019), Initiative D21; Online: https://initiatived21.de/publikationen/d21-digital-index-2018-2019/ (Accessed: 18.06.2019)

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