HADI ASGHARI, NADINE BIRNER, ALJOSCHA BURCHARDT, DANIELA DICKS, JUDITH FAßBENDER, NILS FELDHUS, FREYA HEWETT, VINCENT HOFMANN*, MATTHIAS C. KETTEMANN, WOLFGANG SCHULZ, JUDITH SIMON, JAKOB STOLBERG-LARSEN, THERESA ZÜGER

*Corresponding Author, vincent.hofmann@hiig.de

What to explain when explaining is difficult?

An interdisciplinary primer on XAI and meaningful information in automated decision-making

Explainable AI Research Clinic – Report 2022-01

ABSTRACT

Automated decision making (ADM) systems have become ubiquitous in our everyday lives, enabling new business models and intensifying the datafication of our economies. Yet, the use of these systems entails risks on an individual as well as on a societal level. Explanations of how such systems make decisions (often referred to as explainable AI, or XAI, in the literature) can be considered a promising way to mitigate their negative effects. Explanations of the process and decision of an ADM system can empower users to legally appeal a decision, challenge developers to be aware of the negative side effects of the ADM system during the entire development process, and increase the overall legitimacy of the decision. However, it remains unclear what content an explanation has to include and how the explanation can be made to achieve an actual gain of knowledge for the recipient, especially if that recipient is not an expert. The GDPR provides a legal framework for explaining ADM systems. "Meaningful information about the logic involved" has to be provided to the person affected by the decision. Nonetheless, neither the text of the GDPR itself nor the commentaries on the GDPR provide details on what "meaningful information about the logic involved" precisely is.

This interdisciplinary report discusses this question from a legal, design, and technical perspective. The paper proposes three questions to help formulate a good explanation: *Who* needs to understand what in a given scenario? *What can* be explained about the system in use? *What should* explanations look like in order to be *meaningful* to affected users? The outcomes could potentially not only advance the debate among legal scholars but also help developers and designers to understand the legal obligations when developing or implementing an ADM system.

Legally, the explanation has to enable the user to appeal the decision made by the ADM system. "The logic" can be understood as "the structure and sequence of the data processing". This does not necessarily have to include complete disclosure of the entire technical functioning of the ADM system. Since the explanation is intended to balance the power of the ADM developer with those of the user, this balance has to be at the center of the explanation. The GDPR focuses on individual rather than collective rights. This is the subject of many discussions among scholars. However, the interpretation of the GDPR as protecting mainly individual rights is just the minimum requirement for an explanation. Any explanation going further and also having the protection of collective rights in mind, will be compliant with the GDPR as long as the individual rights are also protected. Therefore, we recommend putting the individual at the centre of the explanation in a first step in order to comply with the GDPR.

With regard to the question "what should explanations look like", we argue that XAI is more than just a technical output. To our view, XAI has to be understood as a complex communication process between human actors and cannot be merely evaluated in terms of technical accuracy. Against this backdrop, evaluating the communication process should accompany evaluating the XAI system's technical performance. Evaluating an explanation created by an ADM system cannot be achieved without involving the user receiving the explanation. Their assessment of what a meaningful explanation needs to entail is an essential prerequisite for XAI. For domain experts, the evaluation of the explanation must include information about potentials, risks, and limitations of ADM systems; explainability starts even before the system is in use and encompasses the complete socio-technical complex.

When it comes to the target group of an explanation, public or community advocates should play a bigger role. These advocate groups support individuals confronted with an automated decision. Their interest will be more in understanding the models and their limitations as a whole instead of only focussing on the result of one individual decision.

As we will demonstrate, there is a gap between how developers and legal experts define what explanations are. Developers aim to debug statements that help them understand their models, but these are less useful for individuals who need explanations to be able to challenge a decision. Also, from a technical perspective,

the term "logic involved" as it is used in the GDPR is – at best – misleading. ADM systems, and data-based systems in particular, are complex and dynamic socio-technical ecosystems. Understanding "the logic" of such diverse systems therefore requires action from different actors and at numerous stages from conception to deployment. Developers have to explain to the ADM system *how* to explain. Methods to explain the explanation often involve using additional approximate models with potentially lower accuracy that raise the question of whether the goal is to explain (justify) the decision or to really understand how the original model arrived at the decision.

Furthermore, transparency at the input level is a core requirement for mitigating potential bias, as post-hoc interpretations are widely perceived as being too problematic to tackle the root cause. The focus should therefore shift to making the underlying rationale, design and development process transparent—documenting the input data as part of the "logic involved". For example, the use of datasheets can lead to more transparency by enabling expert users to better understand the overall process and translate it to lay users. Ultimately, using such measures will help improve ADM systems. In other words, the overall XAI process should involve direct and indirect stakeholders from the very beginning rather than hoping that machine learning models will be able to provide human-compatible explanations post-hoc at the end of the development chain without prohibitive effort.

KEYWORDS

Artificial Intelligence, Automated-decision making, Explainability, Interdisciplinarity, Transparency, Privacy Regulations

CITATION

Asghari, H.; Birner, N.; Burchardt, A.; Dicks, D.; Faßbender, J.; Feldhus, N.; Hewett, F.; Hofmann, V.; Kettemann, Matthias C.; Schulz, W.; Simon, Judith; Stolberg-Larsen, J.; and Züger, T. (2021). *What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making.* Alexander von Humboldt Institute for Internet and Society. https://doi.org/10.5281/zenodo.6375784.

ABOUT THE REPORT

This document is the final report of the "Explainable AI Research Clinic", a five-day impact-driven and interdisciplinary research format focused on specific use cases of explainable AI. In September 2021, this event was hosted by the AI & Society Lab of the Alexander von Humboldt Institute for Internet and Society (HIIG) as well as it was part of the NoC research project "The Ethics of Digitalisation – From Principles to Practises", which aims to develop viable answers to challenges at the intersection of ethics and digitalisation. Innovative formats facilitate interdisciplinary scientific work on application-, and practice-oriented questions and achieve outputs of high societal relevance and impact. Besides the HIIG, the main project partners are the Berkman Klein Center at Harvard University, the Digital Asia Hub, and the Leibniz Institute for Media Research | Hans-Bredow-Institut.

CONTENTS

0 PREFACE – THE STARTING POINT

Artificial Intelligence (AI)[1] is an umbrella term that stands at least for a subfield of computer science, for certain technologies such as expert systems or machine learning, and for an ever growing number of applications such as face recognition or automatic decision support. There has never been an all-encompassing definition of AI, and it will sooner rather than later be worthless to ask if a certain IT system is just digital or already AI; more and more components will be AI-based, including in everyday applications where we will not notice the difference. That said, when we are dealing with (partially) "intelligent" machines, we have certain expectations including getting explanations for the systems' actions. However, the opacity of machine-learning algorithms and the functioning of (deep) neural networks make it difficult to adequately explain how AI systems reach results. They remain a black box. Calls for more insights into how automated decisions are made have increasingly grown louder over the past couple of years. Explainability is therefore the necessary first step in a series of conditions which lead to a decision to be perceived as legitimate: decisions that can be justified are perceived as legitimate. But only what can be questioned can be justified and only what can be justified can be explained. Thus, explainability is a precondition for a decision to be perceived as legitimate (Forst & Günther, 2021).

We need to make sure that we know enough about automated decision making (ADM) systems in order to be able to provide the reasons for a decision to those affected by that same decision – in a way they understand (explainability). Simple enough for it to be understood, yet sufficiently complex so that the AI's complexity is not glossed over. We conceive of ADM systems, and data-based systems in particular, that they are complex and dynamic socio-technical ecosystems, comprised of different actors – individual humans in their various roles (e.g. designers, developers, CEOs), institutions and organisations, as well as technologies (i.e. data, algorithms, models).

Given this complexity, it is not an easy task to ensure that we harness the power of AI for good, and produce explanations on how decisions were reached – albeit being a requirement under European law, such as the General Data Protection Regulation (GDPR). The GDPR aims to protect individual rights when it comes to data protection and digital rights. It can be seen as one of the world's most influential regulations on data protection, privacy and digital rights. This also includes regulation on automated decision making. Despite its standing in the world, many details of how to interpret the GDPR are still unclear.

To tackle the aforementioned challenges, the Research Clinic on Explainable AI explored the meaning of a key phrase in the GDPR on explainable AI: The meaningful information about the logic involved. These legal terms will be explained from a governance, technical and design perspective:

- Governance perspective: What are the requirements regarding explainability in the GDPR and what must be explained in order to meet these requirements?
- Design perspective: What should explanations look like in order to be *meaningful* to affected users?
- Technical perspective: What can be explained about "the logic involved"?

The three perspectives reveal a different understanding of these legal terms. Some of the views may even contradict those of another perspective. However, the aim of this report is not to resolve this conflict in its entirety but rather also bring these contradictions to light for further research. The interdisciplinary approach is intended to form the foundation for understanding the legal terms "meaningful information about the logic involved" of the GDPR, which has hardly been explained so far. The answers of this report to this question is therefore a first, important step towards a better understanding which can root from a combination of the three different perspectives. This report sees itself at the beginning of a research process that hopefully inspires more research to build on it.

[1] In the following, we stick to the widely adopted term XAI. However, rather than using the notion of "AI" consistently, we sometimes also refer to automated decision making (ADM) systems as the debate around explaining an automated decision is also valid for non-AI ADM systems.

1 INTRODUCTION

Three questions can be considered crucial to aid reflection on what constitutes a good explanation: Who needs to understand what in a given scenario? What can be explained about the particular system? How can we integrate the user's view and communicate the system result in a helpful way? The following section will draw on a use case from Finland at various points.

1.1 The loan refusal use case

In 2018, the Finnish National Non-Discrimination and Equality Tribunal prohibited the credit rating practices of the financial company Svea Ekonomi AB. This decision was made upon a charge of the Non-Discrimination Ombudsman of Finland who dealt with the case of a Finnish citizen who was refused a loan. This was surprising for the applicant since he did not have a negative individual credit rating nor any other negative credit records which could explain the refusal. After requesting the reasons for the refusal, the financial company first stated that there was no obligation to justify the decision. Further investigation brought to light that the decision of the company was based on statistical probabilities and the applicant would have been granted a loan if he was counterfactually either a woman, lived in an urban area or if his mother tongue was Swedish rather than Finnish (in Finland). His individual credit rating and income were not relevant factors. The Tribunal prohibited this practice as an act of discrimination with a conditional fine of €100,000 in case of the continuance of the practice.

This case is relevant because it raises a number of key questions:

- What kind of sources (individualised / categories) are ADM tools using?
- Can we explain what sources an ADM system drew on to reach a decision (and on which it did not)?
- How can that knowledge be adequately communicated?
- What is the quality and quantity of information that has to be provided so the person confronted with the decision can understand it (explainability) and accept it (justifiability)?
- Which data points affect the result most?
- Could those have been addressed by the customer?

1.2 What explainability and XAI mean

Generally speaking, an explanation is an answer to a why-question (Ehsan, 2021). According to Krishnan (2020), most definitions of the concepts of interpretability, transparency, and explainability regarding AI-applications are insufficient to solve the problems that they are enlisted for. (These are, for example, knowing how machine learning algorithms work so we are able to regulate them, or knowing why an algorithm makes mistakes, so we can correct them). What in her view is missing, is taking into account that explainability is inherently contextual: the success of an explanation depends on the target group (for example, developers, domain experts, end users, or policy makers), and the purpose of the explanation itself. Therefore, in our understanding, explainability is not a general answer to a why question, but an answer to this why question tailored to be understood by a specific target audience.

In the XAI discourse, explainability is understood in two main directions:

(i) Transparency:
Informally, transparency is the opposite of opacity or blackboxes. It suggests a conscious understanding of the instrument by which the model works. We consider transparency at the level of the entire model, at the level of components and parameters, and at the level of the training algorithm and data. In a wider sense,

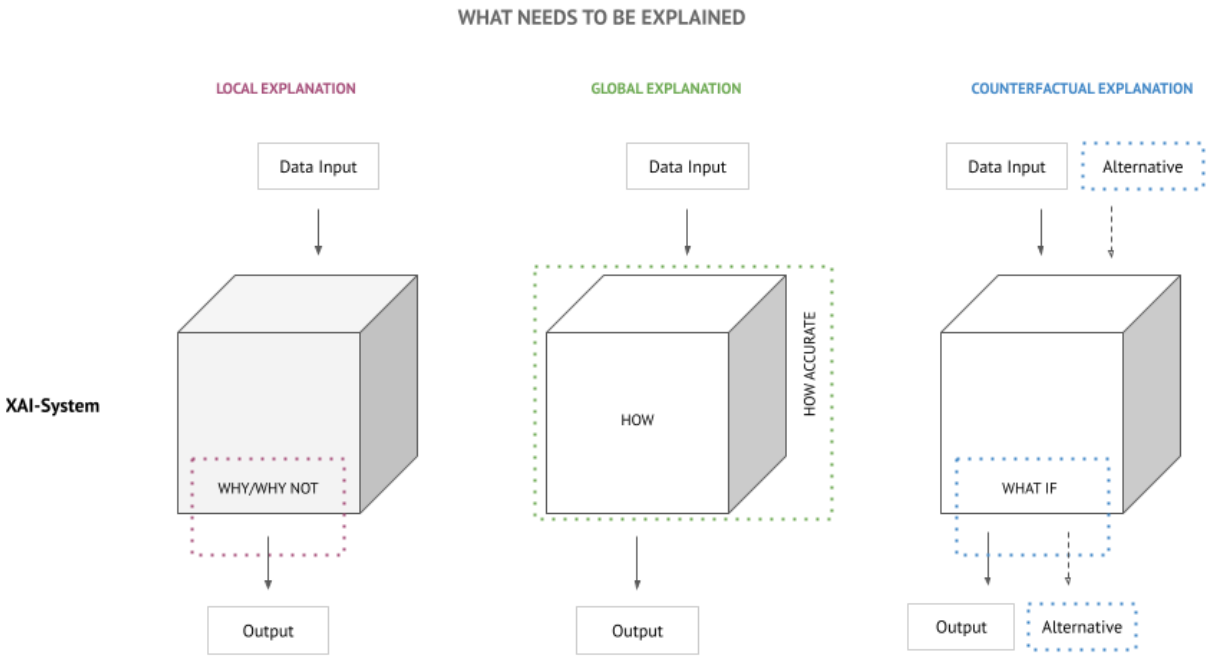
the whole process from designing a system via development and testing up to roll-out and revision might be considered.

(ii) Post-hoc interpretability:
Post-hoc interpretability presents a particular way to deal with extracting information from learned models. While post-hoc understandings do not necessarily provide clarity on how a model functions, they may, in any case, present helpful information for specialists and end clients of machine learning. Some basic ways to deal with post-hoc interpretations incorporate natural language clarifications, visualisations of learned models (heatmaps), and explanations by example.

What we try to build by using explainability methods is: (a) trust, (b) causality chains, (c) transferability, (d) informativeness, (e) ethical foundations.

1.3 What needs to be explained

Regarding XAI, Liao et. al (2020:5f) offer an approach that differentiates distinct components of the answer to the overarching why question in explainable AI (XAI) systems.



Own Illustration on the basis of the explanation approach of Liao et al. (2020)

The components necessary for a comprehensive explanation include:

- What data is used as *input?*
- What is the resulting o*utput data?*
- What is the *performance,* or accuracy, of the system?
- *How* is a decision made in general by this system (*global*)?
- *Why or why not* has a distinct (specific) decision been made? (*local* and factual)?
- *What* outcomes would we have *if* the parameters were different (counterfactuals)?

Besides the question of what to explain, the degree of detail and the format of a representation are crucial. Explanations have to be seen not only in the context of their target group but also in their situational use-context. An explanation given in real-time and where it occurs, might conflict with the useability of a

system. An intuitive use with minimal cognitive effort is generally desired when designing systems, while understanding an explanation (presumably) requires a more conscious mode of thinking. If the context in which the decision is made allows taking time (e.g. a job application), then a complex explanation might be useful; but it can be problematic in situations which demand a quick reaction (Guidotti et al., 2018) (e.g. when driving with an autonomous car). There is, on the other hand, the danger of oversimplification (for the sake of understandability or strategic interests). Thus, finding a fitting explanation method is important.

In general, it is important to note that explanations, if they are not accurate, or if they are badly communicated, could even weaken the user's autonomy. That is why in some cases no explanation about an AI-application's output might be a better solution to empower the user than a misleading explanation.

1.4 Three steps towards a good explanation

In the context of data processing, explainability is discussed in terms of different objectives and, consequently, at different levels of abstraction. It can simply be a matter of better understanding the general functioning of a socio-technical system such as, e.g. a bank lending practice and credit management under certain conditions. In the following, the goal is different and more concrete: an explanation can be a means to an end that a governance system pursues and enables the recipient to exercise their legal rights. For example, it can give a customer the information to recognize that they have been wronged and that they can successfully appeal a decision.

1.4.1 Who needs to understand what in a given scenario?

From this perspective, the first step is to determine who needs exactly what knowledge in order to act in accordance with the governance concept. To stay with the use case example, it does not help a person whose loan application was rejected because of a decision made by an ADM system to obtain more information about the functioning of the system in general (global explanation), which is often the subject of the literature on the rapidly emerging field of XAI. It is more a question of whether a specific decision was made erroneously that has legal relevance (local explanation). Only this kind of information about the system is helpful in the specific context. In the scenario of a medical diagnosis supported by an AI system, a completely different type of information may be relevant for different target groups like the patient, the doctor or an advocate group.

In the research discourse on explainability so far, mainly two target groups are in focus: the developers of a system, in order to gain insights into their own product, and secondly specific groups of expert-users that utilise AI-applications as assistive systems in their field of expertise, such as doctors or researchers (Yang et al. 2018). Critically, explainability for lay users and their representatives (e.g. NGOs that aim to ensure user rights) are typically left out of the picture. If we see the user as a citizen and acknowledge a right to explanation (as also the GDPR requires) the quality of an explanation and its suitability for the general public is central. This research gap, which we want to call 'explainability for the public', will be one focus of this report.

1.4.2 What can be explained about the particular system?

The next step would then be to clarify the question of what statements can be made or generated about the system in question. This depends very much on the technology used, but also on the implementation. There may be information that is already available or that can be generated by tests or audits. It might well be the case that information is required for which the system architecture must be changed or the training data needs to be enhanced in order to generate it. Here, the cooperation of experts with normative expertise (law, ethics) with computer science practitioners is of great importance.

1.4.3 How can we integrate the user's view and communicate the system result in a helpful way?

Different methods exist to integrate the explainability needs of the (expert or non-expert) user (Liao et al., 2020; Wolf, 2019; Eiband et al., 2018). Singh et al. (2021) have researched how to create more suitable explanations for specific user groups and underline that explanations can fail if they do not sufficiently consider human factors. Explanations therefore need to be contextualised regarding their specific user group and domain. Thus, after determining 'who' needs to receive an explanation, and 'why' (that is, what are the explainability goals), the key question becomes 'how?'.

Finally, the fact that the required information is available or can be generated does not mean that the affected person also develops the understanding that allows them to act in the sense of the governance objective, e.g. to legally appeal the decision taken. This process of conveying the available information must be understood as a separate step, as a demanding process of communication that requires its own design, evaluation and monitoring processes.

1.5 XAI in the GDPR

Algorithms crunch data to arrive at models and systems. Data is protected in the EU through the GDPR. The GDPR requires data controllers to provide data subjects with information about the existence of automated decision-making, including profiling[2] and, in certain cases **meaningful information about the logic involved**[3], as well as the significance and the envisaged consequences of such processing for the data subject. Articles 13 and 14 are notification duties imposed on data controllers and Article 15 provides a right to access information throughout processing. Article 22(1) states that a data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or similarly significantly affects them.

An important background towards a better understanding of these norms is the aim of the GDPR and the history of data protection law: It is basically designed to protect individual rights. GDPR empowers data subjects that have been discriminated against but the GDPR does not offer great potential when it comes to protecting group-related and societal interests such as equality, fairness or pluralism (Dreyer & Schulz, 2019). Because Articles 13–15 all relate to the rights of the data subject, meaningful information should be interpreted in relation to the data subject. Or in other words: an individual without any technical expertise must be able to understand why particular information was given. To them, the information about the logic must be meaningful. However, the interpretation of the GDPR protecting mainly individual rights is just the minimum requirement for an explanation. Any explanations going further and also having the protection of collective rights in mind will still be compliant with the GDPR as long as the individual rights are also protected.

The legal community acknowledges the fast-paced dynamic of socio-technical development and does not seek to limit technically useful developments, but advocates instead for a certain flexibility on the technical requirements. "Meaningful information about the logic involved" should not be specified by only naming different types of technology, but by understanding the technical fundamentals and the process of the explanation in general (Selbst & Powles, 2017:233-242).

1.6 Why is an explanation required?

Any societal order, according to Rainer Forst, is related to a "specific understanding of the purpose, the goals and rules of this order" (Forst, 2018:70). The purpose, goals and rules need to be justified. The orders are thus "orders of justification" (:87) and the justifications are formulated as "justification narratives" (:96-97). This becomes particularly virulent when it comes to AI-based decisions. Importantly, everyone has a right to justification of the order in which they exist and the key decisions within that order (Forst,

[2] Articles 13(2)(f), 14(2)(g), and 15(1)(h)

[3] Articles 22(1) and (4)

2017). No one should be subjected to norms and institutions which cannot be justified towards them, based on reasons which they cannot question. Practices of justice are thus based on justifications[4], which are based on justification practices (Kettemann, 2020:109).

2 LEGAL FRAMEWORKS ON XAI

In the following paragraphs, the legal meaning of the "logic involved" and "meaningful information" will be addressed as well as the problem within the GDPR of the "human in the loop".

2.1 What does "logic" mean from a legal perspective?

Article 13 (2) (f), Article. 14 (2) (g) and Article 15 (1) (h) GDPR were introduced to protect the individual rights of data subjects to receive information on the collection of personal data in automated data processing operations. They were designed to ensure a transparent and fair procedure in data processing for the data subject, which is considered necessary by the EU legislator (Dix, 2019). The norms include information obligations for data controllers[5] and therefore play a role in balancing the interests of data subjects and the company running the ADM system by ensuring a transparent and fair procedure in data processing for the data subject (General Data Protection Regulation, 2016). Establishing such a balanced procedure is the aim of the norms which has to be considered when interpreting the terms used in the GDPR like the "logic involved". The term is not legally defined within the GDPR. It is therefore necessary to approach the meaning of the term "logic involved" with tools other than the words of the GDPR itself.

The example of so-called "scoring" like in the Finnish case of loan refusal ("systematic procedures, usually based on mathematical-statistical analysis of empirical values from the past, for forecasting the future behaviour of groups of persons and individuals with certain characteristics"[6] (Verbraucherzentrale Hamburg, n.d.)) makes it clear how broadly the concept of involved logic can be understood: Every form of scoring triggers information obligations about the design and structure of the procedure. This includes information on which collected data (factors) are included in the calculation of the probability value (score value) and with which weighting, and to what extent there is a mutual influence of score values (Schmidt-Wudy, 2021, Art. 34 Rn. 71). According to the German Federal Court of Justice (BGH), complete disclosure of the entire technical functioning of the ADM system like the score formula (scorecard) is generally not necessary, since responsible persons or companies have a legitimate interest in secrecy (BGH 2014, VI ZR 156/13). However, such disclosure is required in cases of the GDPR, provided that the data subject can only notice and have incorrect calculations corrected through this gain in information (Schmidt-Wudy, 2021, Art. 15 Rn. 78.3.). The reason for this is the comprehensibility requirement that applies in particular to information about the logic involved. This states that an "actual gain in knowledge" must be generated for the person concerned which requires comprehensive explanations of the automated decision-making processes (Hoeren & Niehoff, 2018:47).

It is helpful to analyse at what part of an automated decision the different norms of the GDPR require information on the logic involved. Art. 13 (2) (f) focuses on the moment of data collection. Therefore, individual processing results cannot be within the scope of protection of the article (Wachter et al, 2017:76ff). This leads to the conclusion to understand the logic involved as the information related to general "methods and criteria of data processing, such as the functioning of the algorithm used in the creation of a score" rather than certain results (interim or final) of the processing, which is unknown at the

[4] See also 'The Normative Order of the Internet. A Theory of Online Rule and Regulation' (Kettemann, 2020). For comparative perspectives, see 'Deontology of the Digital: The Normative Order of the Internet' (Kettemann, 2020a)

[5] See also Art. 13 (2) (f), 14 (2)(g) and 15 (1) (h) GDPR 2016/679

[6] Translation from German: "systematische, in der Regel auf mathematisch-statistischer Analyse von Erfahrungswerten aus der Vergangenheit basierende Verfahren zur Prognose über das zukünftige Verhalten von Personengruppen und Einzelpersonen mit bestimmten Merkmalen"

time of data collection.[7] The information obligations from Art. 14 (2) (g) GDPR correspond to those of Art. 13 (2) (f) GDPR (Bäcker, 2020 Art. 14 Rn. 18). Therefore, it can be assumed that the term "logic involved" is used identically in these norms. An expansion of the concept of the "logic involved" can be seen in the scope of protection of Art. 15 (1) (h) GDPR. The norm is based on the point in time after the data collection and includes evaluation results and procedural characteristics of the past (Bäcker, 2020 Art. 15 Rn. 27). In this respect, one can speak of a temporal extension in Art. 15 (1) (h). Since the articles use the identical term, this needs to be understood in the broadest way limited by the specific article. Therefore, the logic involved includes both evaluation results and procedural characteristics of the past.

The term "logic involved" is linked to Article 12 (a) of the Data Protection Directive (DPD) (Dix, 2019 Art. 13 Rn. 16). The DPD was the EU legislation on data protection before the GDPR came into effect. The GDPR fully replaced the DPD. However, investigating the meaning of terms used in previous regulations can be helpful to understand the term in the current norms. The German version of Art. 12 (a) DPD guarantees information about the "logical structure of automated processing" (Datenschutzrichtlinie, 2015 Art. 12). In the English version of the DPD, different to the German version, the term "logic involved" is used. Therefore, it can be assumed that the English term "logic involved" consists of the structure and sequence of the data processing (Dix, 2019 Art. 13 Rn. 16).

A clear definition of the logic involved does therefore not exist from a legal perspective. Nevertheless, some concrete requirements could be defined. The interpretation of the term "logic involved" has to align with the overall goal of the GDPR to balance the relation of the data subject and companies running ADM systems. Also, the explanation must include the "logic involved" comprehensively in a way that means an actual gain of knowledge and enables the user to question the decision. A slightly more precise explanation roots in the old DPD, which defines the "logic involved" as the structure and sequence of the data processing. The BGH stated that this does not need to contain a complete disclosure of the score formula if they are not necessary for questioning the decision.

2.2 What is "meaningful information about the logic involved"?

When can an explanation enable the person to legally challenge the automated decision? The concrete scope of the meaningful information to be provided with regard to the "logic involved" – especially of ADM systems – has not been conclusively clarified yet. One could start looking at all laws applicable to an automated decision in general and the specific decision in question. In the example of the Finnish loan use case, these would be the GDPR, anti-discrimination law and Finish national laws regulating automated decision making in general and lending in particular. The downside of this approach is the necessity to constantly adopt the explanation to changes in law, either through legislation or court rulings. It is also challenging to consider all possible laws that an automated decision could potentially be in conflict with to keep the explanation understandable. A minimum solution would be to provide enough information to allow users to understand whether the decisions might be in violation of their rights. It is clear that showing a user, e.g, a graphical representation of a highly complex decision tree model does not make sense. Then again, ADM systems that cannot be explained can also not be used in decisions falling under Article 22 GDPR.

Which information needs to be provided in which way to reach the goal of enabling a person to challenge an automated decision has not yet been satisfactorily answered (Dix, 2019 Art. 13 Rn. 153). This calls even more for a joint attempt at defining the terms utilizing the skills of other disciplines.

[7] Translation from German: "Methoden und Kriterien der Datenverarbeitung, etwa die Funktionsweise des Algorithmus, der bei der Bildung eines Scorewerts genutzt wird": Bäcker, in: Kühling/Buchner, DS-GVO BDSG, Art. 13 Rn. 54; cf. Roßnagel/Nebel/Richter, ZD 10/2015, 455, 458.

2.3 Human in the loop and other obstacles

Besides the already stated difficulties in interpreting the terms of the GDPR, revealing information on the ADM-system comes with other problems. One of them is intellectual property. In the case of a private German credit score rating agency (SCHUFA), courts decided that SCHUFA does not have to reveal information since it is part of their business secret and underlies the intellectual property of the company. Although other cases have been decided differently, intellectual property remains a possible obstacle when demanding insights into automated decisions.

Another problem besides the understanding of the terms of the GDPR is the "human in the loop". With the reference to a right to obtain human intervention, Art. 22 (3) GDPR is based on a widely held understanding that a "human in the loop" is useful or even necessary to secure rights, especially human rights. However, this presupposes two things. First, that humans produce better results from a legal perspective. This is questionable if they, and not the machine, bring prejudices into play. Secondly, the human in the loop has to be allowed to make an independent decision. Here, the debate about "meaningful human control" has not yet been sufficiently reflected in law (Braun et al., 2021:4).
Studies show that at least in certain contexts, people are unlikely to deviate from the proposal of an established ADM system, even if they formally have all the freedom to do so (Meyer & Douer, 2021). This was the case, for example, with employment office staff, where an ADM system made a classification of job seekers. Although the employment office staff were formally allowed to override the automated decision, this happened in only 0.58 % of the cases (Jędrzej & Sztandar-Sztanderska & Szymielewicz, 2015). It seems plausible that meaningful human control requires, in any case, humans to have more information than the result of the ADM system. There is also the question of whether he or she may erroneously expect disadvantages in the event of deviation, such as a doctor who fears to risk liability if he or she performs surgery in deviation from the suggestion of an established system. Other elements of the decision architecture may also play a role.

2.4 Summary

Legal methods reach their limits when it comes to the terms "meaningful information" and "logic involved". The latter could be understood as the structure and sequence of the data processing (global explanation of the system). The information revealed about it has to help balance the relation of the data subject and companies running ADM systems and ensure an actual gain of knowledge to enable the user to appeal the (local) decision. The BGH stated that this does not need to contain a complete disclosure of the score formula if they are not necessary for questioning the decision. What makes information meaningful remains even more obscure.

3 HOW TO DELIVER THE EXPLANATION TO THE RECIPIENT?

After starting with the legal perspective, aspects concerning the design of an explanation will be elaborated upon in the following section.

3.1 Communicating explanations

While different queries from different actors need different types of explanations, as we will further discuss at a later stage, we additionally face the task of communicating explanations in a task and audience-appropriate form. Here, it is essential to assess which kind of epistemic needs can be served best with which measure. For instance: an individual customer demanding an explanation for being refused a loan must be provided with the reasons for this decision and ideally with an explanation of what change in input data would have resulted in a different verdict/output data. Note that here we are looking at what constitutes a good explanation from a design perspective in general, irrespective of what is legally required in a specific situation. In a clinic setting, different types of explanatory interfaces may be needed for expert users of a system like radiologists assessing additional evidence for/against internal bleedings, cancerous

tissue, etc. For lay users like patients getting informed about diagnoses and prospective treatment options, the explanation has to be different.

It is commonly distinguished between local and global explanations – and both need to be provided by the socio-technical ADM system. While the global explanation addresses the system level, the local explanation deals with a concrete decision. Yet other types of visualisations may be needed in auditing contexts where global instead of local explanations are needed. Here as well, different requirements may be posed by experts of different kinds (tech expert/domain expert) and laypeople (e.g. representative groups, NGOs, but also policy/law makers, etc). Given these different task-, system- and audience specific requirements, prototype-testing of both the types and forms of explanations given should be conducted with various user groups.

Going back to the case of the German SCHUFA, as of now, users get very little to no explanation as to why their loan has been refused in a particular case, and also no information is provided about which factors were most relevant in the scoring model overall (global). A relevant explanation in this setting could be given *locally* to a user about an individual loan request, in the form of counterfactuals: which of a user's factors need to change, and by how much, to lead to a different outcome. In the example of the Finish loan refusal, these would be factors like mother tongue and gender. In such a setting, end users of the system have a reasonable need for local explanations, even though legal grounds are not sufficient to guarantee it to be given.

Our line of argumentation here is that explanations of AI are much more than the technical outputs and interfaces: they are a **complex communication process** that starts long before the information output is generated by a system since it also relates to the training data used to build these systems, as well as the justifications for using an ADM system in a particular context. We advocate for understanding this process more holistically as a process taking place between human actors, which might involve technical agents, but the success of a communication can not be evaluated in terms of technical accuracy in the XAI case.
As we will iterate in the following sections, XAI is highly context and case-specific, which makes a one fits all standard nearly impossible. However, we aim to identify, based on existing research and our conceptual analysis, which assumptions and requirements can be formulated for certain groups or similar cases.

3.2 XAI is highly context specific

The frontend design requirements of end users for explanations are highly context specific. This aligns with established (ISO) standards on the usability of interactive systems as "the extent to which a product can be used by specified users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use." (ISO, 2018). To create guidelines (content, format, detail level) on specific explanation requirements within a given domain, scenario-based design / explainability scenarios can be used to make sure user goals and context are taken into account.

Another factor to consider is the degree of AI-assistance involved in the process. It makes a difference if there is limited AI-assistance (e.g. automated lesion measurements) or major AI-assistance (e.g. AI-based computer-aided diagnosis). In contexts of extensive AI-assistance, users typically need to get a much more detailed explanation. This includes results, global as well as local XAI details, to fully understand how AI might have influenced their diagnosis and treatment plan.

3.3 Target groups of AI

We have discussed the fact that explanations are context, task, and target specific. Different targets have different needs, expectations, and existing knowledge which an explanation needs to fulfil and add to (Ribera & Lapedriza, 2019; Liao, Gruen & Miller, 2020; Miller, 2020). This point is generally well understood. Currently, however, many AI explanations are targeted primarily towards developers (who for instance want to be able to better understand their models and debug and improve them). Two other

groups often discussed and understood are the domain experts (e.g. the doctors in a medical setting using AI), and the so-called end or lay user.

We believe there is another target group that is often implicitly talked about, but not made explicit: the public or community advocate. This role is a person or organisation who might represent and follow the interests of a community of users, for instance, an NGO representing cancer patients, online communities of users around topics, as well as civil activists and special interest groups interested in particular topics. These public advocates often have more expertise and ability to understand how AI systems work, and in particular when it comes to issues concerning their community of users, issues related to systematic injustices and discrimination that may originate in the AI system or may originate in society and be exacerbated by the AI systems. These advocates are also often in constant exchange with their community, so they can also explain complex issues around AI and decisions of systems to their community in ways that are understandable for that community.

Assessing and understanding documented information, which is important and effective for the groups at risk, especially if it involves the choice of major medical treatments (e.g. surgery, chemotherapy, etc with possible major side-effects), will be undoubtedly very hard for individuals. It is well known that humans, even highly educated experts, may not be so good at understanding statistics (Kahneman 2013:13), and unpacking metrics and information about the global reliability and accuracy of models is similar. This is where the advocate group can be of service and a key target for explainability.

The advocate group's interest will be more in the global understanding of the models and their limitations. Obviously, the advocates will need to be trained. While one could understand that the involved professionals (for instance doctors) could also be aware of these issues, they might not have the time or necessary sensitivity, which is why such communities and support groups are often formed in the first place. (Also going back to the idea that often one can learn better from peers about complex issues). The advocates' approach, importantly, must not be understood purely in an adversarial manner (towards the development and use of AI systems), but rather also educative and communicative towards the system designers. They can also be good partners for developers in a participatory design process of ADM systems.

Their importance is already understood in other political contexts in society and has been raised in general around issues of inclusive design for AI Systems, i.e. that design teams need more people of colour and women to avoid problems of bias and discrimination (Barocas, Hardt & Narayanan, 2019).

We believe this advocate role should be made explicit and distinct from the professional and individual user. In addition to their time, knowledge and expertise to handle these issues, there is also an important legal reason for this distinction: to recognize them as a separate target for an explanation, which would need kinds of information about a system that other individuals might not need. In this regard, additional legislation might be necessary, since for instance the GDPR's Art. 22 looks at explanations individually. By understanding the needs of the advocate role, the level of access to information and documents necessary for the advocates to be able to function, their role can be better articulated.

Target group / Question	Developer / Researcher	Domain experts	Advocate	Individual user
Why	verification, improvement, accountability	learning how to use the system in context, awareness for risks and limitations	to establish grounds to advocate for vulnerable groups and issues	enable informed decisions, self-advocacy, empowerment

Target group / Question	Developer / Researcher	Domain experts	Advocate	Individual user
What	global model, data representation why and why not	independent global training	– ideally global explanation. Minimum: datasheet, documentation about bias testing – answers to issue related inquiries, – provided with local explanation by individual user	– local explanation, – access to global explanation should be possible → could become an advocate
When	throughout the whole AI-lifecycle	– previous to using the system, – during application	on demand, also it could be even before the system is in action in a participatory design process	real time or post hoc, close to when the decision is made
How	*context dependent*			
	intrinsic	visualisation, natural language	documentation with technical facts and natural language	conversation questions, why not
Evaluation	completeness test, performance	critical survey of trust → whether the system is trustworthy, rather than asking if people trust the system	– completeness, comprehensiveness, absence of discrimination / fairness – quality of explanation → active role in process	satisfaction questionnaires, experiments

Table 1: Different needs of different target groups. Adapted from Ribera and Lapedriza (2019) with additional target group and content.

3.4 Explanation Moment: training for domain experts, training regarding limitations, and documentation

Explainability starts even before the system is in use. When understanding the act of explaining automated decisions as a communication process, we have to acknowledge that the process starts before the actual use of an AI system. Further, this means it becomes important in which moment explanations are given, which is again dependent on task, context and target group. Also, education is part of the explanation: Before the actual employment of an AI system, target groups, especially domain experts, have to be educated about the risks, limitations and capability of AI systems they intend to use. In addition to this, general knowledge of the potentials and limitations of AI in their field is necessary. An important question is who can provide such information in a non-biased manner. For the use of AI in radiology, medical professional associations could be suitable providers for such a training course. Such associations might at the same time function as an example for the target group of the advocate. It should be further investigated if general education on AI and specific use-contexts generally fall into the competencies of advocate groups as well.

4 WHAT IS THE "LOGIC INVOLVED" FROM A TECHNICAL PERSPECTIVE?

After the legal and design perspectives, the following section will look from a technical perspective at which forms of "logic" exist, what can be explained and what cannot be explained, and how to deal with forms of logic that are difficult to explain.

4.1 Which types of logic exist?

From a technical perspective, the term "logic involved" is – at best – misleading. At first sight, it seems to be linked to symbolic AI (e.g., knowledge-based systems such as rule-based expert systems) rather than subsymbolic AI (e.g., statistical machine learning). This is puzzling as the former is precisely the type of AI that is more or less fully transparent. The black-box problem is much more present in the subsymbolic realm, as can be seen in today's artificial neural networks. The elaboration we presented above in 2.1 about the logic involved in "data processing" can be interpreted so that this refers to the training of a model, which would be more appropriate. Yet the terms "algorithm" and "creating a score" again sound like the idea of someone devising (programming) a "formula" that can be fed with data and then, e.g., calculates a score that grants a loan or not. If this is the case, this view is much too naive to be helpful in tackling the problems at hand.

We conceive of ADM systems, and data-based systems in particular, that they are complex and dynamic socio-technical ecosystems, comprised of different actors – individual humans in their various roles (e.g. designers, developers, CEOs), institutions and organisations, as well as technologies (i.e. data, algorithms, models). In addition to this socio-technical distribution of agency, such systems are also characterised through their temporality. Different actors are involved in different stages from conception, design and development to deployment and the systems may also change over time (Zweig et al., 2018). Understanding "the logic" of such diverse systems therefore requires action from numerous actors and at numerous stages from conception to deployment.

For now, we will abstract away from this socio-technical dimension and will turn to the systems in a narrower sense. Both complexity and dynamics of technical systems can differ profoundly, ranging from simple, linear models over Support Vector Machines to Machine Learning (deep learning) models all the way to (the relatively few) systems which continuously learn and adapt their behaviour in the wild.

4.2 What can and what cannot be explained?

As mentioned before, it is commonly distinguished between local and global explanations and it depends on the target group if a local or global explanation is more suitable. While the global explanation operates at the system level and targets the overall behaviour, the local explanation deals with a concrete decision. The requirements of what is needed to provide "meaningful information about the logic involved" may comprise of technical or organisational mechanisms to make *local or global* explanations.

From a technical point of view, we can generally distinguish three different types of ADM systems to which XAI can be applied:

1. Symbolic reasoning: Explainability is high, performance is usually very much limited. Expert systems, for example, are still being used in industry, finance, etc. These systems lend themselves to problems that can be formulated with grammars, logical formulae, etc., which is not the case in many real-life applications.
2. Classical Machine Learning using Bayesian statistics and feature engineering: Some models like decision trees lend themselves comparably well for human inspection, others require more efforts to become interpretable.

3. Machine Learning with neural networks (deep learning): This technology is the current state-of-the-art in many areas and brings the black box problems widely known from the public debate about AI. This is the context of the current XAI debate. The performance is usually better than when using classical Machine Learning (ML).

In principle, everything can be explained if the system has access to the knowledge needed. Yet in many cases, it will be prohibitively expensive and complex to build the respective systems that can provide explanations on a high level of abstraction and comprehensibility. One concrete example is the use of AI systems to support radiologists in detecting tumours on MRI scans: if we train a neural network with MRI scans and for every image there is a label (tumour/no tumour), we may end up with a model that can predict a tumour in a new image with a high probability. If we want it to predict the region, then we need to mark the region in the training images, if we want medical explanations, then we need to annotate the training images with medical information of the right kind, and we probably need to annotate many more training images than for the original task, as we need to make sure that all possible explanations occur in the training data often enough. What often makes the intransparent systems so efficient is the fact that we don't need to prepare specific data for them. We also echo the argument made by Krishnan (2020) that the advantage of uninterpretable systems lies exactly in their uninterpretable nature: these systems have a high performance, because they can often detect patterns in a way that is different to the functioning of the human brain. This is not intuitive to a human and therefore, an explanation of how these models operate is often not intuitive as well.

4.3 What to do if there is no easy explanation?

Many existing AI systems depend on models which automatically learn feature representations and use an intransparent internal functionality (e.g. neural networks). Despite recent attempts to better examine such systems (Samek et al., 2021), the trustworthiness of such examinations are still debatable and their usefulness could mainly be empirically established (Jacovi et al., 2021:624ff; Ehsan & Riedl, 2021). Although it holds true that everything can theoretically be explained if the system has had access to the necessary information, there is often a discrepancy between what the system has had access to and what a target group might expect from a valuable explanation; the black-box problem.

One option that is sometimes chosen is to provide post-hoc explanations. These interpretations might explain predictions without elucidating the mechanisms by which the model works. The model is fixed in this setting and a prediction has been made by the model. After this, hence "post-hoc", an explainability method acts upon this and provides an explanation. One example of post-hoc interpretations is natural language explanations which provide a (free-text) rationale about the model prediction based on additional information, e.g., from an external "explanation model". A more widespread example are saliency maps (or heatmaps) used to analyse deep neural networks. They highlight the parts (or features) of the input (image, text, tabular data, etc.) to the model that are deemed most important (or salient) to the model prediction according to the model. This is then presented to the user as a coloured layer on top of the data. Traditionally, red colours correspond to the most salient features while blue colours are features indicating against the prediction. Yet, for high stake use cases simply providing post-hoc "model examinations" is still problematic as the user does not receive comprehensive information about the "logic involved" in a way that could enable them to appeal a decision. Inherently interpretable models, as Rudin (2019) argues, are faithful to what the model actually computes, while post-hoc explainability methods that "act from the outside" are not.

The focus should instead be on appropriately balancing the inherent interpretability of the model and the significance of the model output. We propose to make the design and development process of the ADM system as well as the underlying rationale transparent. We conceive this to be the more promising approach than only focusing on inner-technological means in shedding some light into the black box of ML through heatmaps or other types of automatically produced visual or verbal partial examinations of the systems' behaviour. This may require obligations to document the processes of data gathering and preparation

including annotation or labelling. The method selection for the main model like architectural choices as well as the extent of testing and deployment should also be documented. Such documentation should be obligatory for any ADM systems dealing with personal data as well as other high-impact systems.

Data documentation

One method to explain an ADM-system without only using post-hoc explanation could be transparent data documentation. Datasheets have been recently proposed as a method to describe the characteristics of datasets. Gebru et al. argue that documenting "motivation, composition, collection process, [and] recommended uses" (2021:2) would "facilitate better communication between dataset creators and dataset consumers" (:1). In the context of XAI, datasheets mitigate potential biases by ensuring that datasets are used in suitable contexts (Bender & Friedmann, 2018; Gebru et al., 2021). Ensuring the highest amount of transparency at this level – the input level – is arguably one of the best ways to influence and truly understand what is happening in a model. By fully documenting what exactly is in the dataset, we can be aware of datasets that do not represent certain members of society, for example. Datasheets could lead to limiting the data used for training models to datasets that can be described by humans. This would rule out extremely large models, where even the developers do not know what exactly is in the input data. The data statements proposed by Gebru et al. (2021) and Bender & Friedman (2018) are written in a way that they can at least be understood by expert users: the statements should ideally contain as much information as possible about the data but do not explicitly state whether the data is suited (or not) for a particular use-case. This implies that these expert users can then explain the characteristics of the data to lay users and make a judgement on the suitability of the data for the specific application.[8] Jacovi et al. (2021) argue that for a formalisation of human-AI trust, reproducibility checklists and fairness checklists are relevant as well for defining contracts.

Architectural choices

As stated above, we understand an explanation as part of an ongoing socio-technical process. This makes the future explanation a relevant factor for decisions about the design of the backend (technical functioning) and frontend (which the user sees and interacts with). An important aspect when providing transparency on architectural choices is the justification for using large, complicated models: why have we decided to use a so-called black box algorithm here?[9]

The balance between interpretability and performance needs to be weighed up, particularly in high-stake decisions. As stated above (in Sec 4.2), intuitive explanations can often not be provided by uninterpretable systems, which are however still being used, as their advantage lies in their high performance. However, this high performance is also due to the fact that they are based on procedures that are not humanly intuitive and therefore difficult for humans to understand.
Irrespective of the legal requirements (where an explanation is in our interpretation almost always a prerequisite), at a minimum for high-stake decisions[10], it may be favourable to use a more interpretable[11] model, i.e. not a deep neural model. When a neural model, however, offers supreme performance, it may be worth tweaking the architecture in such a way that the model does not offer a definitive decision or

[8] The actual metrics used to evaluate a system also need to be conveyed accurately in order to establish a level of trust with the expert and lay users, i.e. what accuracy did a certain system have on a certain task. This level of transparency is also highly relevant when multiple software tools are used together—which is very often the case. Having a standardised way of documenting these details would ensure that various tools are combined in ways that ensure that the whole software chain is transparent.

[9] We use ever increasing amounts of compute in our neural networks because of the way these state-of-the-art models scale (Kaplan et al., 2020)

[10] For example in contexts where explanations are needed to make informed decisions, or to allow advocate groups to help prevent biases or injustices.

[11] Every model is interpretable, but not every explanation has to be meaningful to the user. Meaningful information can come in many forms as Gilpin et al. (2018) write about features of explanation methods: (1) Type of the problem faced; (2) Explanatory capability used to open the black box; (3) Type of black box that can be explained; (4) Type of input data provided to the black box.

classification, but instead gives the user suggestions (while ensuring that the user can meaningfully control what to do with the system output). This would be a possible solution although it should be kept in mind that just formally leaving the human to make a choice does not mean that the human will divert from the ADM system's decision like described earlier. While this does not resolve the issue of understanding the inner workings of the AI system, end decisions will rely on expert user interpretations of the assistive model output. As such, final decisions can be explained by the human rationale of the expert user for which the system is designed.

4.4 How does the technical understanding differ from the legal understanding

We think that the demand for documenting the input data, architectural choices and evaluation methods falls under what was called the "structure and sequence of the data processing" in the legal considerations above. We propose to make the design and development process of the ADM system as well as the underlying rationale transparent. This may require obligations to document the processes of data gathering and preparation including annotation or labelling. Ensuring the highest amount of transparency at the input level is arguably one of the best ways to influence and truly understand what is happening in a model. Likewise, instantiating the demand of showing the "logic involved" by the more concrete requirement of providing local or global explanations of the socio-technical systems can be seen as an operalisation step.
To sum up, the current technical understanding of explanations is more geared towards developers whereas the legal understanding is more focused on the individual affected by the decision. From a technical perspective, we also see the necessity to weigh up the risks and benefits of non-interpretable decisions instead of consistently banning them.

5 CONCLUSION

The report examined the meaning of two legal terms of the GDPR, the "**meaningful information** about **the logic involved**." Legal methods reach their limits when it comes to clear definitions of these terms. As we argued here, the logic involved could be understood as the structure and sequence of the data processing. The information revealed about it has to help balance the relation of the data subject and companies running ADM systems and ensure an actual gain of knowledge to enable the user to appeal the decision. The BGH stated that this does not need to contain a complete disclosure of the score formula if they are not necessary for questioning the decision. What makes information meaningful remains even more obscure.

Looking at the terms from a design perspective leads to understanding an explanation more generally: Explainability starts even before the system is in use. When understanding the act of explaining automated decisions as a communication process, we have to acknowledge that the process starts before the actual use of an AI system. Further, this means it becomes important in which moment explanations are given, which is again dependent on task, context and target group. Also, education is part of the explanation: Before the actual employment of an AI system, target groups, especially domain experts, have to be educated about the risks, limitations and capability of AI systems they intend to use. Further, general knowledge of the potentials and limitations of AI in their field is necessary. It is important to identify who can provide such education in a non-biased manner. For the use of AI in radiology, medical professional associations could be suitable providers for such education. Other associations, which focus on patient rights, might be seen as an example for the target group of the advocate. As we argued, general education on AI and specific use-contexts generally can benefit from the involvement of advocates, such as NGOs or representatives of special interest groups.

Technically seen, the demand for documenting the input data, architectural choices and evaluation methods falls under what was called the "structure and sequence of the data processing" in the legal considerations above. Likewise, instantiating the demand of showing the "logic involved" by the more concrete requirement of providing local and/or global explanations of the socio-technical systems can be seen as an operalisation step. Summing up, we are convinced that the discussion in this document has found a good common vocabulary for turning the legal imprecision and the technical imperfection into actionable

frameworks to make the design and development process of the ADM system, as well as the underlying rationale, transparent. This could ensure that the current systems can be inspected and future systems might become more elaborate.

Understanding an explanation as an ongoing process, evaluating, and adapting the explanation is essential. When evaluating a model, the context is obviously highly important: how do we know when a system is good enough? The developing team needs to document this decision in a transparent way. In the best case, this decision should not be made solely by a team of developers. Instead, a communicative process should be utilised to fully understand all possible requirements and implications of a system. This process should engage a variety of people as wide as possible: stakeholders, expert users, lay users, relevant advocates.

The evaluation of XAI is a research field which is slowly gaining more attention. Following our previous argument, that explanations for AI are a complex communication process that involves human actors at the core, this evaluation can only be successfully achieved if the whole process is looked at – instead of only the technical elements of the process. Whether the explanation process is "successful" can only partly be determined by auditing the AI systems' functionality. Another important aspect of this evaluation is to ask the recipients of an explanation, which in our examples are multiple different target groups, to which degree the explanations served its purpose. Obtaining feedback from the recipients of the statement must be done in the context of research projects that can use the responses to determine the effectiveness of the statement for the respective target groups.

REFERENCES

Bäcker, M. (2020) in Kühling, J., & Buchner, B. (Ed). *Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO / BDSG*: Vol. XXII (3rd ed.). München: C.H. Beck.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning. Limitations and Opportunities*. fairmlbook.org. https://fairmlbook.org/

Barocas, S., & Selbst, A. D. (2016). *Big Data's Disparate Impact*. *SSRN Journal*, *104*. https://doi.org/10.2139/ssrn.2477899

Bender, E. M., & Friedman, B. (2018). *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.* Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). P*rimer on an ethics of AI-based decision support systems in the clinic*. J Med Ethics, 47(12). https://doi.org/10.1136/medethics-2019-105860

Dix, A. (2019) in Hornung, G., & Spiecker, Simitis. S. (Ed.), *Datenschutzrecht*. Baden-Baden: NomosKommentar.

Datenschutzrichtlinie (2015). 'Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr'. Available at: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A31995L0046

Douer, N., & Meyer, J. (2021). T*heoretical, Measured, and Subjective Responsibility in Aided Decision Making.* ACM Trans. Interact. Intell. Syst., 11(1), 1–37. https://doi.org/10.1145/3425732

Dreyer, S., & Schulz, W. (2019). *Künstliche Intelligenz, Intermediäre und Öffentlichkeit. Bericht an das BAKOM*. https://www.bakom.admin.ch/dam/bakom/de/dokumente/bakom/elektronische_medien/Zahlen%20und%20Fakten/Studien/bericht-chancen-risiken-intermediaere-2020.pdf.download.pdf/Bericht_Chancen_Risiken_Intermedia%CC%88re_310720_fin.pdf

Ehsan, U., & Riedl, M. O. (2021). *Explainability Pitfalls: Beyond Dark Patterns in Explainable AI.* Human-Computer Interaction.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). *Bringing Transparency Design into Practice.* 23rd International Conference on Intelligent User Interfaces, 211–223. https://doi.org/10.1145/3172944.3172961

Friedman, B., & Nissenbaum, H. (2017). *Bias in Computer Systems. 14*(3), 330–347. https://doi.org/10.4324/9781315259697-23

Forst, R. (2017). *Kritik der Rechtfertigungsverhältnisse. Perspektiven einer kritischen Theorie der Politik* (1st ed.). Suhrkamp.

Forst, R. (2018). *Normativität und Macht. Zur Analyse sozialer Rechtfertigungsordnungen*. Suhrkamp, Berlin.

Forst, R. (2021). *Normative Ordnungen* (G. Klaus, Ed.). Suhrkamp.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). *Datasheets for datasets. Commun.* ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

General Data Protection Regulation (2016). 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)'. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining Explanations: An Overview of Interpretability of Machine Learning*. https://doi.org/10.1109/dsaa.2018.00018

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR), 51*(5), 1-42.

Hoeren, T., & Niehoff, M. (2018). *KI und Datenschutz – Begründungserfordernisse automatisierter Entscheidungen*. RW, 9(1), 47–66. https://doi.org/10.5771/1868-8098-2018-1-47

International Organization for Standardization. (2018). Retrieved January 25, 2022, from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Jędrzej, N., Sztandar-Sztanderska, K., & Szymielewicz, K. (2015). *Profiling the Unemployed in Poland: Societal and Political Implications of Algorithmic Decision Making*. https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). *Formalizing Trust in Artificial Intelligence*. 624–635. https://doi.org/10.1145/3442188.3445923

Kahneman, D. (2013). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. https://arxiv.org/abs/2001.08361

Kettemann, M. C. (2020). *Normative Order of the Internet*. Oxford University Press, USA.

Kettemann, M. C. (2020a). *Deontology of the Digital: The Normative Order of the Internet*. In M. C. Kettemann (Ed.), Navigating Normative Orders (pp. 76–91). Campus.

Krishnan, M. (2020). *Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning.* Philosophy & Technology, 33(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Liao, Q. V., Gruen, D., & Miller, S. (2020). *Questioning the AI: Informing Design Practices for Explainable AI User Experiences.* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3313831.3376590

Lipton, Z. C. (2016). *The Mythos of Model Interpretability.* ArXiv:1606.03490 [Cs, Stat]. http://arxiv.org/abs/1606.03490

Miller, T. (2019). *Explanation in artificial intelligence: Insights from the social sciences.* Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Ribera, M., & Lapedriza, À. (2019). *Can we do better explanations? A proposal of user-centered explainable AI.* IUI Workshops.

Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). *Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications.* Proceedings of the IEEE, 109(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Schmidt-Wudy, F. (2021) in Brink, S., Wullf, H., *BeckOK Datenschutzrecht.* 38. Edition, C.H. Beck, München.

Selbst, A. D., & Powles, J. (2017). *Meaningful information and the right to explanation.* 7(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Singh, R., Ehsan, U., Cheong, M., Riedl, M. O., & Miller, T. (2021). *LEx: A Framework for Operationalising Layers of Machine Learning Explanations.* ArXiv:2104.09612 [Cs]. http://arxiv.org/abs/2104.09612

Verbraucherzentrale Hamburg, *Was ist Scoring*. (n.d.). Retrieved January 25, 2022, from https://www.vzhh.de/themen/finanzen/was-ist-scoring.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *SSRN Journal,* 7(2), 76–99. https://doi.org/10.2139/ssrn.2903469

Wolf, C. T. (2019). *Explainability scenarios: Towards scenario-based XAI design.* Proceedings of the 24th International Conference on Intelligent User Interfaces, 252–257. https://doi.org/10.1145/3301275.3302317

Yang, Y., Yan L. F., Zhang X., Han Y., Nan H. Y., Yu C. H. (2018). Glioma Grading on Conventional MR Images: A Deep Learning Study with Transfer Learning. *Front Neurosci. 12*(804). 10.3389/fnins.2018.00804

Zweig, K. A., Fischer, S., & Lischka, K. (2018). *Wo Maschinen irren können: Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung.* https://doi.org/10.11586/2018006