Artificial Intelligence in Medical Diagnoses and the Right to Explanation

*Thomas Hoeren and Maurice Niehoff**

Artificial intelligence and automation is also finding its way into the healthcare sector with some systems even claiming to deliver better results than human physicians. However, the increasing automation of medical decision-making is also accompanied by problems, as the question of how the relationship of trust between physicians and patients can be maintained or how decisions can be verified. This is where the right to explanation comes into play, which is enshrined in the General Data Protection Regulation (GDPR). This article explains how the right is derived from the GDPR and how it should be established.

Keywords: Data Protection, Privacy, AI, Articial Intelligence, Algorithm

I. Introduction

Initial scientific research on artificial intelligence (AI) dates to the 1940s.[1] Since then, technical development has made rapid progress. AI is becoming more and more important in recent years, due to the rapidly increasing computing power and the emergence of huge amounts of data, the big data.[2] Especially AI in the healthcare sector is subject to a strong rise, it is even the fastest developing AI-market with a total volume of almost $700 million and an expected growth up to $8 billion by 2022.[3] AI offers several opportunities for faster, more precise and effective medical diagnoses and decisions. Due to big data databases AI is capable of analysing a huge scope of input data, such as biometric data, external factors, family history, personal medical records and literature.[4] As an example, the Watson Oncology Advisor, made by IBM, is using almost 15 million pages of medical literature to advise oncologists on cancer diagnoses and chemotherapy-plans[5], with an accuracy of 90%.[6] The AI-advice is becoming more and more relevant to physicians' decisions.

Beside the positive aspects of AI-based decisions, one must consider the special topicality due to sensitivity of data used. This is followed up by the debate of the introduction of the General Data Protection Regulation (GDPR), which is dealing with sensitive data and automated decisions made by AI.

This article focuses on these automated decisions using algorithms and AI in the healthcare sector, ar-

DOI: 10.21552/edpl/2018/3/9

[*] Prof Dr Thomas Hoeren is Director of the Civil Law Department of the Institute for Information, Telecommunications and Media Law (ITM) at the Westfälische Wilhelms-University Münster, Germany. Ass Iur Maurice Niehoff is a research assistant at the Institute for Information, Telecommunications and media Law (ITM) at the Westfälische Wilhelms-University Münster, Germany.

[1] Christian Honey, 'Künstliche Intelligenz - Die Suche nach dem Babelfisch' *Zeit Online* (23 September 2016) <http://www.zeit.de/digital/internet/2016-08/kuenstliche-intelligenz-geschichte-neuronale-netze-deep-learning> accessed 27 June 2018.

[2] Wolfgang Hoffmann-Riem, 'Verhaltenssteuerung durch Algorithmen - Eine Herausforderung für das Recht' (2017) 142 Archiv des öffentlichen Rechts, 6.

[3] J Braun, 'Dr. KI, zur Visite bitte!' *Süddeutsche Zeitung* (13 September 2017) <http://www.sueddeutsche.de/wirtschaft/gesundheitssystem-dr-ki-zur-visite-bitte-1.3665374> accessed 27 June 2018.

[4] Joaquin Sarrion Esteve, 'Health Data Treatment. An Approach to the International and EU Legal Framework' in R Arnold, R Cippitani and V Colcelli (eds), *Genetic Information and Individual Rights* (vol 1, University Regensburg 2018 Series Law & Science) 36-53.

[5] Thilo Weichert, 'Big Data im Gesundheitsbereich' (ABIDA report, 2018) 1, 24 <http://www.abida.de/sites/default/files/ABIDA%20Gutachten-Gesundheitsbereich.pdf> accessed 27 June 2018; IBM, 'Watson Health promotion video' <https://www.ibm.com/watson/health/oncology-and-genomics/oncology/> accessed 27 June 2018.

[6] In breast cancer cases, ASCO Post, 'SABCS 2016: IBM Watson for Oncology Platform Shows High Degree of Concordance with Physician Recommendations' (13 December 2016) <http://www.ascopost.com/News/44214> accessed 27 June 2018.

gues whether there is a right to explanation of the data subjects and what requirements must be met.

In a first step, the article defines the relevant terms regarding AI, automated decisions and health data and states its potential risks and challenges (II.1.). Afterwards it argues that there is a right to explanation in the GDPR to overcome these risks (II.2.), followed by an elaboration of the requirements these explanations must fulfil (II.3.). The article exposes what exactly an 'explanation' is and the different stages of an explanation while establishing the reference to health sector decisions and how these explanations can be technically implemented. The article ends with the conclusion that a paradigm shift is taking place in healthcare sector and the right of explanation is an important component to deal with it (III.).

II. Artificial Intelligence in Medical Diagnoses

As already mentioned, on the one hand AI offers great opportunities for a more precise and faster, therefore more effective way to support physicians' diagnoses and decisions.

Examples include a cardiac examination that executes a cardiac segmentation in 15 seconds instead of 30 minutes[7] and a system that assesses breast cancer mammography results 30 times faster than a human physician with an error rate of 1%.[8] This leads to a relief of the physicians, because crowded doctor's offices and hospitals could be reduced. This in turn leads to a better care of the patients. With the support of AI-systems, physicians are able to concentrate more on the contact with patients.

AI is even capable of replacing human decision-making. Using its deep neural networks, and thus assessing huge amounts of medical data, it is as precise as human physicians and often even more accurate. For instance, the already mentioned Watson Oncology Advisor with an accuracy of 90% and the mammography result evaluator with a 1% error rate. Particularly patients with rare diseases are still frequently asking doctor after doctor, obtaining different isolated diagnoses from each of the doctors.[9] This is where a data-based supported AI-system could be of great value to quickly find the right diagnose.

On the other hand, AI-diagnoses and -decisions hold several risks such as privacy and bias.

AI-systems are only as good as their databases, so it must be ensured that the data used is verified and unbiased. The increasing importance of AI-decisions in healthcare could lead to biased decisions as well as the effect that human physicians completely rely on the AI[10], without questioning it critically, although it only show correlation, not causality.

A physician needs to know if and why she can trust an AI-decision and hence she needs a possibility to verify these decisions.[11]

In addition to that, AI-decisions based on complex neural networks are not fully comprehensible, especially not for patients. Consequently, the relationship between patient and physician, which should be based on trust, could be disturbed.[12]

Another challenge, which comes along with the black box decisions, is the task of verifying and contesting AI-results.

The article hereto argues that there is a right to explanation in the GDPR to balance the interest of using the advantages of AI-systems and their risks. The right to explanation relates to the regulation of automated decision-making (Article 22 GDPR).

1. Definitions

a. Algorithms and Artificial Intelligence in the Healthcare Sector

Algorithms and AI are widely used in the healthcare sector.

[7] 'Künstliche Intelligenz in der Medizin: Arztunterstützend, nicht arztersetzend' (*Ärzteblatt.de*, 21 November 2017) <https://www.aerzteblatt.de/nachrichten/83587/Kuenstliche-Intelligenz-in-der-Medizin-Arztunterstuetzend-nicht-arztersetzend> accessed 27 June 2018.

[8] PWC, 'Künstliche Intelligenz revolutioniert die Medizin' (25 July 2017) <https://www.pwc.de/de/gesundheitswesen-und-pharma/kuenstliche-intelligenz-revolutioniert-die-medizin.html> accessed 27 June 2018.

[9] Ärzteblatt.de, 'Medizin' (n 7).

[10] 'over-reliance'. Adrian Bussone et al, 'The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems' (Conference Paper International Conference on Health Informatics, Dallas, USA, October 2015) 1 <http://dx.doi.org/10.1109/ICHI.2015.26> accessed 27 June 2018.

[11] 'Every far-reaching decision should be made accessible for appropriate validation by a human expert' Wojciech Samek et. al, 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models' [2017] ITU Journal: ICT Discoveries, Special issue no 1, 1, 2.

[12] W Nicholson Price II, 'Black-Box Medicine' (2015) 28(2) Harvard Journal of Law & Technology 425.

i. Algorithms[13]

Algorithms are used to systematically solve a problem. They work with the help of deterministic, stringently following, unambiguous and finite rules of action. The input of a certain value is followed by the output of a result, whereby the same result is always obtained due to the determinism with the same input values. Classical example in the analogue world is a cooking recipe, where a clear sequence of actions (recipe) is always followed by the same result (finished dish), it is an 'if..., then...-process'.[14] In the digital world, the rules of action are represented and processed by computer programs. These simple if-then-algorithms were used in medical science years ago, however they were not able to learn, but just supported physicians by executing their if-than-defaults.[15] This means they remain evidence-based and stick to the particular scientific knowledge, are thus validated and ensure a better alignment to the individual patient and therefore improve efficiency.[16] The advantage is that they are still explainable and understandable.[17]

One example is the 'Heart Disease Predictive Instrument', which advises the physician based on 50 predefined factors through a simple yes-no-survey, whether a patient with chest pain should be sent to the coronary care unit or the regular nursing bed.[18]

ii. Artificial Intelligence[19]

AI is also based on the algorithms described above.

Artificially intelligent applications also make use of rules of action but go far beyond that. AI is generally[20] referred to as AI when algorithms achieve the ability to simulate human action.[21] In order to achieve the most human-like action possible, so-called artificial 'neural networks' are created. These correspond to the structure of the human brain. A neural network consists of input and output neurons and intermediate layers, the so-called 'hidden layers'.[22]

This construction is particularly capable of machine learning and its sub-area of deep learning. It includes, in addition to the linear if..., then...-process, the possibility of self-learning ability.

Where pure machine learning is based on the ability to learn by means of human influence, the system learns contexts without any human intervention in deep learning. The system is trained using big data components, ie large amounts of data. Based on the training data, the system recognizes correlations, structures, questions the initial result and improves itself.[23]

In the healthcare sector, the system is fed with millions of input data, ie health records or medical literature, even wearable tech records. With big computing power, the system assesses this huge amount of data almost immediately and explores patterns and correlations. This kind of work cannot be executed by human physicians in terms of time and scope.[24]

The data used is classified as 'data concerning health' (Articles 9(1) and 4(15) GDPR (see b.)). This learning process leads to an increase in the layers between the input and output neurons, enabling increasingly complex decisions.[25]

A famous example is a neural network which was trained with almost 130,000 images of skin cancer to visually detect cancer. The performance was as accurate as a human dermatologist.[26]

Furthermore, the major innovation is that AI is able to explore even more complex non-evidence-based patterns and correlations. This implies AI does

[13] A definition can be found at Armin P Barth, 'Algorithmik für Einsteiger' (2nd edn, Springer 2013) 8; DFK Bitkom, 'Künstliche Intelligenz, Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung' (2017) 1, 67.

[14] Barth, 'Algorithmik' (n 13) 2.

[15] Braun, 'Dr. KI' (n 3).

[16] Price II, 'Medicine' (n 12) 465-466 calls it '(explicit) personalised medicine'.

[17] ibid.

[18] Julian N Marewski and Gerd Gigerenzer, 'Heuristic decision making in medicine' (2012) 14 Dialogues in clinical neuroscience 77-89.

[19] Christian Ernst, 'Algorithmische Entscheidungsfindung und personenbezogene Daten' (2017) Juristische Zeitung 1026, 1027; Wolfgang Ertel, 'Grundkurs künstliche Intelligenz' (4th edn, Springer 2016) 1; DFK Bitkom, 'Künstliche Intelligenz' (n 13) 28-31.

[20] There still is no scientific consensus on a definition, see DFK BitKom, 'Künstliche Intelligenz' (n 13) 28-31 or also Ertel, 'Grundkurs' (n 19) 1.

[21] DFK BitKom, 'Künstliche Intelligenz' (n 13) 28.

[22] Yann LeCun et al, 'Deep Learning' (2017) 521 Nature Deep Review 436, 437.

[23] Jürgen Schmidhuber, 'Deep learning in neural networks: An overview' (2015) 61 Neural Networks 85, 86.

[24] W Nicholson Price II, 'Artificial Intelligence in Health Care: Applications and Legal Issues' (2017) 14 The SciTech Lawyer, 1.

[25] Itamar Arel et al, 'Deep Machine Learning - A New Frontier in Artificial Intelligence Research' (2010) November IEEE Computational Intelligence Magazine 15.

[26] Andre Esteva et al, 'Dermatologist-level classification of skin cancer with deep neural networks' (2017) 542 nature international journal of science, 115 <https://www.nature.com/articles/nature21056> accessed 27 June 2018.

not stick to simple pre-determined scientific knowledge, but explores new patterns, which might not even be understandable by the physicians.[27]

For example, AI might find a correlation, that on middle-aged female patients who are smokers and are diagnosed with bipolar disorder, medication xy is working especially well.

It mostly is no longer possible for a human being to completely understand how an AI-result was achieved - we know that it works without knowing how it works.[28] The decision basis, the original algorithm, is also subject to constant change. The decision becomes an opaque black box[29] for the data subjects.

b. Data Concerning Health in the GDPR

Health data are, in respect to their high level of intimacy, particularly sensitive. First protection measures date back to Hippocrates (460 - 370 BC), whose Hippocratic Oath can be considered as the original safeguard to patients' privacy.[30] This type of medical confidentiality is now standardised in the code of medical ethics and criminal law, eg Section 203 German Criminal Code (*Strafgesetzbuch*, or StGB).[31]

The necessity for data protection rights regarding to health data is derived from Articles 2, 3, 8, 35 EU Charter of Fundamental Rights (EUCFR) that shows that these sensitive data are linked to privacy.[32] This consideration was laid down by the GDPR in Article 9(1).

According to Article 9(1) GDPR, health data are special categories of personal data and therefore higher demands must be made on processing them. Health data, or 'data concerning health', as the GDPR calls it, are legally defined in Article 4(15) GDPR. Thus, data concerning health means 'personal data related to the physical or mental health of a natural person, including the provision of healthcare services, which reveal information about her health status'. The definition is specified by Recital 35 GDPR. There is uncertainty in scholarship, about the scope of the term 'data concerning health', especially data from private apps and wearables, so-called 'quasi-health' data.[33] This article will not expand this aspect, but sticks to the principle of an extensive interpretation to protect the data subjects, whose basis was developed by the *Lindquist* case verdict of the Court of Justice of the European Union (CJEU).[34]

c. Requirements for Automated Decisions in Article 22 GDPR

AI shows its specific strength through automated decisions, meaning the AI reaches a self-acting choice that affects a human being, without human interaction and that is much faster and more effective than human decision-making. As an example, the mentioned Oncology Advisor is capable of deciding about chemotherapy medication of patients on his own.

However Article 22(1) GDPR prohibits those automated decisions.

Article 22(1) GDPR prohibits the data subjects from being a 'subject to a decision based solely on automated processing [...] which produces legal effects [...] or similar significantly affects'.

i. 'solely'

The first important requirement is a decision 'based solely on automated processing'.

This means a procedure that is carried out without human intervention from the acquisition of the data to the issue of the decision.[35] This raises the question of when a decision is considered to be solely automated.[36]

This article distinguishes between four phases of influences of human being in automated processing.

First, a decision-making process carried out from beginning to end without any human influence or oversight, such as adopting a chemotherapy medica-

[27] Price II, 'Medicine' (n 12) 432-433.

[28] Oliver Stiemerling, 'Künstliche Intelligenz - Automatisierung geistiger Arbeit, Big Data und das Internet der Dinge' (2015) 12 Computer & Recht 762, 764; Ertel, 'Grundkurs' (n 19) 308-310.

[29] Frank Pasquale, 'The black box society' (Harvard University Press 2015) coined that term; Price II, 'Medicine' (n 12) coined the term 'Black-Box Medicine'.

[30] Weichert, ABIDA (n 5) 10.

[31] ibid.

[32] J Sarrion Esteve, 'Treatment' (n 26) 5.

[33] For further information see, Gianclaudio Malgieri and Giovanni Comande, 'Sensitive-by-distance, Quasi Health Data in the Algorithmic Era' (2017) 3 Information, Communication and Technology Law.

[34] Case C-101-01 *Bodil Lindqvist* [2003] ECLI:EU:C:2003:596. The court has ruled that information about a broken foot and a partial doctor's certificate on a private website are data concerning health.

[35] GDPR, recital 71 states explicitly so.

[36] Mario Martini in Paal and Pauly (eds), *Datenschutz-Grundverordnung* (CH Beck 2018) art 22 ref 16-18.

tion decision of the Oncology Advisor by the human physician without scrutinising it, is an automated process.

Second, it is unclear whether Article 22 GDPR also covers those processes in which the AI completely prepares a decision, but in which a human ultimately implements the decision without wanting to influence the decisions content ('nominal human involvement'[37]). This is the case with a mere confirmation of the result.[38] One must speak of an automated decision at least, if the 'human in the loop' has no competence to change the decision and it therefore is just a 'token gesture', such as the nurse instead of the doctor in charge confirming the result.[39]

Third, the Guidelines of the Article 29 Working Party claim that a it is necessary to have a human in the loop, whose 'oversight of the decision is meaningful' and who has the competence to change the decision, such as the doctor in charge.[40] Inversely this means, the Working Party allows it to suffice that the human in the loop with competences just oversees and scrutinises an AI-decision.

This article demands higher standards of influence of the human in the loop. As a fourth phase, it claims a real, meaningful influence of the human being. There must be an actual modification of the decision, not just a critical oversight.[41]

In this respect, the oversight or mere decision-making (pressing the 'OK'-button) of the human being in the third phase is not to be taken into consideration. This would ultimately render the standard useless. Also, human intervention in the neural network to improve decisions, such as supervised learning[42], does not constitute sufficient human action. It has no influence on the content but is comparable to maintenance. It must therefore be based on whether the person who is involved in the decision-making process also deals with the content of the decision. This argument goes beyond mere consent.[43]

This can be derived from the purpose of Article 22 GDPR. The purpose of the prohibition in Article 22(1) GDPR is to protect the data subjects from an exclusively computer-based decision. At the end of every decision there must be a human being.[44] The background to this are the fundamental rights protected under Articles 7 and 8 EUCFR as well as national legislation, eg Article 2(1) of the German Constitution and Article 2(1) in conjunction with Article 1 of the German Constitution, the general freedom of action and the right to informational self-determination. For the data subject, it must remain transparent whether she has been the target of a fully automated decision, otherwise a 'feeling of helplessness'[45] towards this decision arises. Furthermore, an exclusively algorithm-based decision concerns the identity and right of self-determination of each data subject. The algorithm processes the acquired personal data based on predefined criteria and weighings, draws conclusions and contexts from them and achieves a result. The data subject is nothing more than a collection of data input, the individual personality of the patient is not taken into account.[46] At latest this would hold the risk of the over-relying effect of decision support systems, which means the human physician will place too much weight on the AI-decision instead of relying on his own knowledge.[47]

ii. 'legal effect' / 'similar significant effect'

Another relevant requirement of Article 22(1) GDPR is that automated decisions must produce 'legal effects' or at least 'similar significant effects'. Although many aspects are controversial here as well, one can leave these issues out. Medical decisions or diagnoses, such as the mentioned chemotherapy medication decision, mostly have a legal effect (Articles 1-3 EUCFR: human dignity, right to life, right to the integrity of the person). At least those decisions will regularly have similar significant effects on the data subjects due to the sensitivity of the data concerning health, which will have an impact on the identity and right of self-determination.[48]

[37] Lilian Edwards and Michael Vaele, 'Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for' (2017) 16 Duke L & Tech Rev 18, 45.

[38] Hoffmann-Riem, 'Verhaltenssteuerung' (n 2) 36.

[39] Article 29 Working Party (A29WP), 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (2017) WP 251, 21.

[40] ibid 21.

[41] Martini in Paal and Pauly (n 36) 19.

[42] Hereto and Schmidhuber, 'Neural Networks' (n 23) 89-91; Stiemerling, 'Automatisierung' (n 28) 763.

[43] Martini in Paal and Pauly (n 36) 17.

[44] Mario Martini, 'Algorithmen als Herausforderung für die Rechtsordnung' (2017) Juristische Zeitung 1017, 1019.

[45] Ernst, 'Entscheidungsfindung' (n 19) 1030.

[46] ibid.

[47] Edwards and Vaele, 'Slave' (n 37) 45.

[48] Martini in Paal and Pauly (n 36) 27.

The interim result is that Article 22 GDPR has a wide scope and therefore the data subjects need to receive sufficient safeguards, such as the right to explanation.

2. A Right to Explanation in the GDPR

The development of AI in the healthcare sector obviously has its benefits, as already mentioned above. Nevertheless, one should not lose sight of potential risks and challenges. Biased decisions, over-reliance in the AI-decision, comprehensibility, verification and contesting the AI- results are the keywords.

The article argues that there is a right of explanation for data subjects (patients) in the GDPR.

a. Article 22(1) GDPR

Article 22(1) GDPR sets the course for this consideration. It could include both, either a right of the data subject or a general prohibition of solely automated decision. There are four main arguments for classifying paragraph 1 as a right.

The precise wording of paragraph 1 speaks of a '*right* not be subject' and it is systematically categorised as a '*right* of the data subject' in Chapter 3.[49] Moreover paragraph 4 clearly speaks of a prohibition ('shall not'), so one can argue the lawmaker uses a different wording in paragraph 1 to emphasise that it is precisely a right rather than a prohibition.[50] This could finally be strengthened by the historical interpretation that even Article 15 Data Protection Directive 1995 wording was a 'right'.[51]

Despite these strong arguments, this article considers that paragraph 1 contains a prohibition.

The wording 'right' in the context of Article 22(1) GDPR cannot be seen as a typical 'right'. A right usually consists of the enablement to actively do some kind of action. In contrast, Article 22(1) GDPR just enables a passive right not to be subject to the decision.

Systematically, paragraph 1 is built like a prohibition compared to paragraph 2; these two norms are in a rule-exception ratio. Paragraph 2 represents a typical exception from paragraph 1, 'paragraph 1 shall not apply'. Therefore, it only makes sense to classify paragraph 1 as the general prohibition.[52] Furthermore, connecting to the systematic argument, paragraph 1 opens the possibility to execute connected rights of Articles 13-15 and 22(3) GDPR and is not a right itself. Especially looking at the consequences of the decision whether to classify Article 22(1) GDPR as a prohibition or a right, one should prefer the prohibition. According to this, data controllers (physicians) must fulfil the requirements of Article 22(2) GDPR exception (consent or necessity to perform a contract), otherwise they are acting unlawfully. Respecting data concerning health, even higher requirements are set. Article 22(4) GDPR demands the requirements of Article 9(2) lit. a) or g) GDPR, consent or public interests, in combination with suitable measures to safeguard the data subject's rights.

Added to this, Article 22(2) GDPR comes with the safeguard of Article 22(3) GDPR (see below). If Article 22(1) GDPR is classified as a right, there would be no legal consequence until the data subject executes her right.[53] So only a prohibition offers sufficient protection to the data subjects by granting the connected safeguards automatically.

In practical terms, this means a physician who wants to apply automated decisions like the Watson Oncology Advisor's chemotherapy plan has to fulfil the requirements of an exception of Article 22(2) GDPR as well as ensure suitable measures to safeguard the patients' rights. Recording obligations could be considered here. Otherwise, this processing would be prohibited.

This article argues for a prohibition as a first safeguard to data subjects, but anyway, Article 22(1) GDPR does not contain a right of explanation.

b. Article 13(2)(f) and Article 14(2)(g) GDPR

A right to explanation also cannot be derived from Article 13(2)(f) or Article 14(2)(g) GDPR. Connecting factor could be the wording 'meaningful information about the logic involved', whereby information about the logic involved could mean an explanation of the

[49] Isak Mendoza and Lee A Bygrave, 'The Right not to be Subject to Automated Decisions based on Profiling' in Tatiani Synodinou et al (eds), *EU Internet Law: Regulation and Enforcement* (University of Oslo Faculty of Law Research Paper No 2017-20, Springer 2017, Forthcoming) 9-11.

[50] ibid.

[51] Sandra Wachter et al, 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation' (2017) International Data Privacy Law 38 <http://dx.doi.org/10.2139/ssrn.2903469> accessed 27 June 2018.

[52] Mendoza and Bygrave (n 49) 9; A29WP, 'WP 251' (n 39) 23.

[53] Wachter et al, 'right to explanation' (n 51) 39.

explicit decision.[54] However a 'timeline problem'[55] occurs.

Article 13(1) GDPR explicit wording 'at the time when personal data are obtained' shows the requirement to support the data subject with meaningful information temporally before or during the decision-making. Hence, it is only logical that the data controller must and can only provide the data subject with information about the general basic mechanisms[56], not the specific decision, which is not made yet. [57]

This ex ante explanation is less an explanation than a mere informing of the existence of an automated decision (as the article will emphasise in the next chapter). It is rather suitable to make a data subject aware of the data processing in the first place than to offer a ground to an effective possibility to contest a decision.[58]

So, Article 13(2)(f) and Article 14(2)(g) GDPR contain a notification duty about informing of the existence of automated decision, not a right to explanation as this article claims. Anyhow, it constitutes a second safeguard, meaning that physicians in any case have a duty to inform the patients about the usage of an automated decision before the decision was made and to inform about its general basic mechanisms.

c. Article 22(3) GDPR and Recital 71

As mentioned above, Article 22(3) GDPR applies by reference of Article 22(1), Article 22(2) and Article 22(4) GDPR. Paragraph 3 contains three explicit safeguards: the right to obtain human intervention, the right to express her point of view and the right to contest the decision.

As one can see, the wording itself does not contain an explicit right to explanation. It rather enumerates the minimum requirements of data controllers' safeguards, speaking of the term 'at least'.[59]

This is where Recital 71 comes into play. Recital 71 is the only passage in the whole GPDR which explicitly states a right to explanation, saying a suitable safeguard should include the right to 'obtain an explanation of the decision reached after such assessment'.[60] So this could be the remedy. Only the legal nature of recitals raises questions. Recitals are, unlike the main article text, not legally binding, otherwise, they would have been placed in the main body.[61] That does not mean they do not have any value at all. Recitals, especially Recital 71, are often given too little importance in scholarship. Recitals show the purpose of the legislator, they are guidance to interpret the GDPR.[62]

By arguing there would be no right to explanation because the legislator shifted the right of explanation from the body text of the European Parliament's Report GDPR draft to the recital of the GDPR[63], scholars give too little importance to recitals. One cannot conclude from the fact that the right to explanation is set in the recital and not in the main body that the right is not intended.[64] The fact that the legislator placed an explicit right to explanation in Recital 71 shows the intention the legislator had. If the legislator had not wanted a right to explanation at all, he would not have written it in the recital either. There is no need for the legislator to explicitly regulate every conceivable case. Law consists of undefined legal concepts, which are accessible to the interpretation of courts. The undefined legal concept of Article 22(3) GDPR ('suitable measures', 'at least') is therefore qualified by Recital 71 and its right to explanation. This goes along with the main goals of the GDPR, to strengthen data protection and the rights of data subjects.[65] Article 22(3) GDPR must be broadly interpreted because it is teleologicly intended by Recital 71. Otherwise the explicit safeguard of paragraph 3, con-

[54] Bryce Goodman and Seth Flaxman, 'European Union regulations on algorithmic decision-making and a "right to explanation"' (2016) 6 < https://arxiv.org/abs/1606.08813> accessed 27 June 2018.

[55] Wachter et al, 'right to explanation' (n 51) 15.

[56] ibid 6 are speaking of 'system functionality'.

[57] Sandra Wachter et al, 'Counterfactual Explanantions' (2018) 31 Harvard Journal of Law & Technology 38, 44 <https://ssrn.com/abstract=3063289> accessed 27 June 2018; Gianclaudio Malgieri and Giovanni Comande, 'Why a right to Legibility of Automated Decision-Making Exists in the GDPR' (2017) 3(3) International Data Privacy Law 1, 19.

[58] Malgieri and Comande, 'Legibility' (n 57).

[59] Wachter et al, 'right to explanation' (n 51) 10.

[60] ibid 9.

[61] Among others, Mendoza and Bygrave, 'Profiling' (n 49) 8, 16; Edwards and Veale, 'Slave' (n 37) 49; Malgieri and Comande, 'Legibility' (n 57) 19.

[62] Tadas Klimas and Jurate Vaiciukaite, 'The Law of Recitals in European Community Legislation' (2008) 15 ILSA Journal of International & Comparative Law, 7.

[63] Wachter et al, 'right to explanation' (n 51) 12.

[64] Wachter et al, 'Counterfactual' (n 57) 40 are explicitly claiming that.

[65] Andrew D Selbst and Julia Powles, 'Meaningful Information and the Right to Explanation' (2017) 7 (4) International Data Privacy Law 233, 19.

testing a decision, cannot effectively be executed. The data subjects must be capable of effectively enforcing their right to contest. This is made possible by the right of explanation.

d. Article 15(1)(h) GDPR

A right to an ex post explanation of a specific decision can also be derived from Article 15(1)(h) GDPR. Although it has the same wording as Article 13(2)(f) and Article 14(2)(g) GDPR, there is no 'timeline-problem' in this article. The executing of this right depends on the active action of the data subjects. Hence, it is possible to execute the right both before (ex ante) and after the decision was made (ex post).[66] It also contains a right to explain the specific decision, and not just the general system functionality. Even though the wording of Article 15 GDPR seems to be future orientated like the Articles 13 and 14 GDPR ('envisaged consequences', 'existence')[67], again one must realise the purpose of the GDPR, instead of clinging to overly narrow interpretations of words. The strengthening of data protection and data subject rights leads to a broad interpretation to guarantee effective safeguards for data subjects. Thus, Article 15(1)(h) GDPRs requirements of 'meaningful information' must be interpreted in such a manner that data subjects are capable to effectively contest automated decisions. One must focus on the word 'meaningful'. To ensure this, 'meaningful information' stands for a preferably high level of comprehensibility, which can be achieved best by explaining the specific decision, not just general system functionality (see next chapter).[68]

As a further interim conclusion, every automated decision that fulfils the requirements of Article 22(1) GDPR must match paragraph 2, otherwise it has to face the penalties of the GDPR. Furthermore, Article 13(2)(f) and 14(2)(g) GDPR contain the duty of the data controller (physician) to inform the data subject (patient) about the existence of an automated process and its general system functionality from an ex-ante point of view. Finally, derived from Article 22(3) with Recital 71 and Article 15 (1)(h) GDPR, a right to explanation of a specific automated decision from an ex post point of view exists.

Altogether this triumvirate of prohibition, notification duty and right to explanation can offer an optimal and comprehensive protection for the data subjects.

3. Explanation Requirements

After elaborating the right of explanation, the next step is to evaluate how to shape the ex post right to explanation of a specific decision and thereby the explanation itself. An effective enforcement of rights is already grounded in Articles 6, 13 European Charter of Human Rights (ECHR) and Article 47 CFREU.[69] One must keep in mind the purpose of this safeguard, to challenge potential risks of automated decisions in the healthcare sector. Hence the explanations must be shaped in such a way that the patient as a data subject is able to comprehend, verify and most of all effectively contest automated decisions. The overall aim is to strengthen the trust in the relationship between patient and physician and not to deliver them to the automated decision, which could lead to a feeling of being at the mercy of the AI.[70]

The article first lists the statutory requirements of the GDPR and defines a specific proposal in a further step.

a. Article 12(1) GDPR Requirements

Article 12(1) GDPR sets basic requirements to informing duties of the data controller, which are also useful to specify the requirements for explanations.

Paragraph 1 states measures to provide the data subject with information that are 'concise, transparent, intelligible and easily accessible form, using clear and plain language'. These requirements are repeated by Recital 58 sentence 1. Recital 58 sentence 3 recognises the conflict the data subjects have, namely that 'the technological complexity of practice makes it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected'. Furthermore, Recital 63 sentence 1 sets the

[66] Edwards and Vaele, 'Slave' (n 37) 52; Wachter et al, 'right to explanation' (n 51) 17.

[67] Wachter et al, 'right to explanation' (n 51) 17-18 cite this as one of the main reasons to deny a right to explanation in art 15 GDPR.

[68] Malgieri and Comande, 'Legibility' (n 57) 23.

[69] Wachter et al, 'right to explanation' (n 51) 32.

[70] Ernst, 'Entscheidungsfindung' (n 19) 1030.

requirement that the data subject should have the possibility to 'verify the lawfulness of the processing'. By way of example, Recital 63 sentence 2 declares an access of patients to medical records and diagnoses.

From this compilation the intention of the GDPR can be drawn. The GDPR intends to make the data subject capable of understanding a decision autonomously,[71] meaning there must be a minimum level of comprehensibility for common human beings.

b. What Does 'Explanation' Mean?

Proceeding from this general principle, the article proposes the following requirements to an explanation of automated decisions.

i. Model-Based Explanation of System Functionality (MCE)[72]

The article does not follow the idea of model-based/model-centric (MCE) explanations. These explanations open up the black box.[73] In order to understand an automated decision, the whole model of an AI-system, the system's functionality, shall be disclosed. This can comprise the algorithm models, decision trees, training data, parameters, weighings, categories and source codes. A complete disclosure of the whole model promises a comprehensive insight into the functioning of the AI-system to completely understand a decision-making. However, this approach is subject to several hurdles that cannot be overcome.[74]

First, there would be a legal hurdle. Opening the black box and disclosing the system functionality would face and affect the rights of the AI-developer as data controllers.[75] Data controllers' interests of trade secrets and intellectual property are affected. For example, the functionality of the Oncology Advisor by IBM has an enormous business value that IBM wants to protect. This is taken up by Recital 63 sentence 5 which states a 'right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property'. However, Recital 63 sentence 6 declares that 'the result of those considerations should not be a refusal to provide all information to the data subject.' It amounts to a balancing of the interests involved. In the case of disclosure of the model-based explanation, the rights of the data controllers will regularly outweigh. For one thing, model-based disclosure would lead to an obstacle to innovation, because it would render the effort of data controllers worthless. The particular market value of developed AI-systems, especially on the basis of specific expertise, must be protected against copying and gaming its system. Another point is that disclosure of the whole model is neither necessary nor suitable to the explanation requirements, because there are technical and contentual hurdles as well.

A technical hurdle exists because AI-models, particularly neural networks, are not easy to disclose. AI-systems corresponding to neural networks are not programmed according to a linear model of a line code but continue to program themselves. This means that the model is self-developing, it learns by itself. Therefore, it is not possible to perform ordinary linear control. Usually even the developers themselves do not know how the AI-system works and how it is making its decision.[76] The developers only know that it works. The model could not have been published at all, as it is constantly evolving itself, at least it would be a huge effort to do so.

Lately, this kind of disclosure would be of no use to the data subjects in content-wise terms either. A disclosure of the whole AI-model, eg providing the algorithm model, is not understandable for a common data subject and therefore not useful. For instance, a patient will not be capable of comprehending and contesting a decision by disclosure of mathematical formulas in bits and bytes. The patient needs an explanation that is interpretable for common human beings.[77] The parallel to human decision-making can be drawn here. Patients also do not need and do not want to know how a human physi-

[71] Malgieri and Comande, 'Legibility' (n 57) 2.

[72] Among others, Edwards and Veale, 'Slave' (n 37) 56.

[73] 'Decompositional Explanation', see Gregoire Montavon et al, 'Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition' (2015) 1 <https://doi.org/10.1016/j.patcog.2016.11.008> accessed 27 June 2018.

[74] Wachter et al, 'Counterfactual Explanations' (n 57) 3 are speaking of 'four barriers', but with different kind of content.

[75] See Malgieri and Comande, 'Legibility' (n 57) 30-32 and Wachter et al, 'right to explanation' (n 51) 24-26 for more details on the discussion.

[76] Among others, Joschua A Kroll et al, 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review 633, 638.

[77] Finale Doshi-Velez and Mason Kortz, 'Accountability of AI Under the Law: The Role of Explanation' (Havard Public Law Working Paper No 18-07, 2017) 1, 2, 6-8 <https://ssrn.com/abstract=3064761> accessed 27 June 2018.

cian diagnose was made in the brain, how her neurons worked.[78] Patients rather need to know why the human physician made a particular decision or diagnose, meaning which factors and thoughts lead to the specific decision (see ii.).

Anyhow, the effort to disclose the entire model of the AI-system is disproportionate to the benefit for the data subjects, because there is barely a benefit.[79] Therefore, after weighing the interests of data controllers and data subjects a disclosure of the AI-model is not necessary. There is neither a need for a physician to explain the whole AI-system she uses, nor could this help patients to understand a decision.

This article argues to use a more specific disclosure, the subject-based disclosure of a specific decision.

ii. Subject-Based Explanation of the Specific Decision (SCE)[80]

Following the idea of a parallel to human decision-making, the article argues that a subject-based explanation of a specific decision suits best the requirements of explanations. To make an automated decision understandable, comprehensible and therefore contestable for common human beings, the extremely complex neural network decision-making of AI-systems has to be broken down into factors understandable for humans.[81] These factors comprise the main factors of the decision as well as their relevance.

An example is a patient diagnosed with a 95% risk of getting diabetes via a complex AI-system, working with a neural network and huge data amount. She must be presented with comprehensible factors such as 'weight', 'genetic predisposition' or 'nutrition' to understand the origins of the diagnose 'diabetes-risk'. Furthermore, she also needs to know the significance and impact of the single factors. The different factors of a diagnose do not all have the same impact, some are more, some are less influential to the diagnose.[82] You can find these two key requirements in human decision-making as well. A patient will expect the physician to explain her decision-making process by revealing the key factors that lead to her decision.[83] It is in the nature of things that to understand a particular decision, one must understand the key factors that lead to it, rather than simply believing the decision.[84] This is a key requirement for a self-determined patient and must also apply to automated decision-making. A data controller must not hide behind an AI-system.

This kind of explanation can overcome the legal hurdle of trade secrets and IP rights of data controllers. Since it does not comprise disclosure of internal logic of AI-models, (disclosure of source code, algorithm model etc) this kind of explanation intervenes much less in the rights of data controllers. Here the interests of the data subject to an explanation and the rights of the data controller are likely balanced. The data subject has a legitimate interest in explaining automated decisions but does not receive a complete insight into the black box, she rather gets the main factors of decision-making, all the same as regarding to a human decision-making.[85]

In the end it is also technically realisable. Unlike disclosing the whole AI-model and opening the black box, disclosing just the main factors and its weighing of a decision is clearly more feasible. The solution is to make AI-decision-making explainable, not to explain the AI-model.[86] For instance, this is technically possible with so-called Local Interpretable Model-Agnostic Explanations (LIME). Through a technical procedure, the relevant word fields around the decision are recognised ('local'). It is not capable to explain or interpret a whole AI-model. It only recognises the local, neuronal activities around the specific decision. In the event of a special diagnose e.g. 'diabetes-risk', the factors 'overweight', 'genetic

[78] ibid; Dmtry Larko, 'Explaining the model or Making black box transparent' (2016) 12 <https://github.com/h2oai/h2o-meetups/blob/master/2016_11_28_UC_Berkeley_DeCal/2016_11_28_UC_Berkeley_Model_Explanation.pdf> accessed 27 June 2018: 'We don't need to understand how a brain works to understand why a person made a particular decision'. Also Edwards and Veale, 'Slave' (n 37) 43: 'we often do not understand how things in the real world work'.

[79] One of the few higher court verdicts of national courts in Germany also ruled that the trade secrets of the data controllers outweigh the information rights of the data subjects, the *SCHUFA* judgment, judgment of the German Federal Court Bundesgerichtshof 28 January 204 – BGH NJW 2014, 1235. The verdict based on the dated Data Protection Directive 95/46/EC that was implemented in German national law in BDSG-Alt.

[80] Among others, Edwards and Veale, 'Slave' (n 37) 56.

[81] Doshi-Velez and Kortz, 'Accountability' (n 78) 8.

[82] ibid.

[83] ibid 9-10, 12.

[84] Wachter et al, 'Counterfactual Explanations' (n 57) 8.

[85] Peter Schantz in Schantz and Wolff (eds), *Das neue Datenschutzrecht* (CH Beck 2017) 745 comes to the same conclusion in relation to these requirements.

[86] JA Kroll et al, 'Algorithms' (n 77) 650-652; C Seifert et al, 'Visualizations of Deep Neural Networks in Computer Vision: A Survey' in Tania Cerquitelli, Daniele Quercia and Frank Pasquale (eds), *Transparent Data Mining for Big and Small Data* (Springer 2014) 123, 123-125.

predisposition' and 'hypertension' might be identified as relevant for the result.[87] These factors are therefore simple enough and comprehensible for layperson humans like patients ('interpretable'). The patient is able to interpret the diagnose by means of his expectations and knowledge. By reference to the diagnose and the understandable factors, the patient is capable to comprehend the decision-making process in a self-determined manner and to use this explanation to assess whether the decision is based on factors that are correct and appropriate or not[88], just like a physician would explain a human decision-making to her. Here too, the physician would list the corresponding factors that lead the physician to her decision-making and a patient could logically understand the decision, as far as this is at least possible as a layperson.

Another positive and preferable effect would be that physicians have to deal with the comprehensibility of automated decisions as well. Legally, they are data controllers according to the GDPR, but have no insight into the AI black box either. They also need this explanation of how the AI-result was made by disclosing the relevant factors of the AI-decision. Based on their expert knowledge, physicians can critically question the AI-result, make a diagnosis and explain it to the patient.[89] Without this explanation physicians would only be operator of automated AI-decisions. So the role of physicians is changing, from classical evidence-based medical diagnosticians to an advisor and controller of AI-decisions.[90].

In addition, a patient needs to know which factors have what kind of impact on the decision-making process. Factors like 'overweight' and 'hypertension' might have a distinctly higher impact on a diabetes-risk diagnoses than 'gender' or 'average sleep duration'.

To forge a bridge to human decision-making process again: In human decision-making diagnoses physicians would also explain which factors are particularly important and which are more likely to be secondary factors.

This is made possible by 'counterfactual explanation'. Counterfactual explanation is working the way that it explains a specific decision by describing how it is changing if someone modifies a particular factor of the decision.[91] So if a patient is diagnosed a 95% risk of getting diabetes, she will be wondering how she can modify (decrease) this risk. Therefore, one can feed the AI-system with modified factors, such as a decreased factor of weight, to see how the diagnose is changing. By trying out and modifying factors, the various impacts of the factors on the diagnose can be identified. For instance, statements can be made such as: 'If your BMI was 24 instead of 28, your risk of getting diabetes would only be 68%.'[92]

Therefore, it can be summarised that a combination of LIME and counterfactual explanation fits the requirements of the right to explanation of a specific decision.

III. Paradigm Shift

Due to the progressive technological development AI-technology is becoming more and more important. AI-systems are already playing a consequential and useful role in decision supporting and development goes in such a way that it will also find large fields of application in decision-making.

One can say a paradigm shift is taking place in the healthcare sector.[93]

The raise of automation could lead to a change of the relationship between human being and machine.[94] Physicians will take on new roles. One of their tasks will be of an advisory and controlling nature. It will be important how to deal with this shift.

On the one hand, AI-systems in the healthcare sector are of great value and extremely useful, so there should be no obstacles through over-regulation.[95] On the other hand, a paradigm shift must not lead to a loss of trust in the relationship of physician and pa-

[87] Marco Tulio Ribeiro et al, '"Why Should I Trust You?" Explaining the Predictions of Any Classifier' (2016) 2 et seq <https://arxiv.org/abs/1602.04938> accessed 27 June 2018.

[88] Patrick Hall et al, 'ideas on interpreting machine learning' *O'Reilly* (15 March 2017) <https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning> accessed 27 June 2018.

[89] Ribeiro et al, 'Trust' (n 88) 2.

[90] 'Artificial Intelligence changes Evidence Based Medicine'(Scalable Health White Paper, 2017) <https://www.scalablehealth.com/Resources/WP/AI_Changes_Evidence_Based_Medicine.pdf> accessed 25 August 2018.

[91] Wachter et al, 'Counterfactual Explanations' (n 57) 5-7.

[92] ibid 21.

[93] Jennifer Lepies, 'Künstliche Intelligenz in der Medizin: "Wir wollen Ärzte nicht arbeitslos Machen"' *Heise-Online* (7 September 2017) <https://www.heise.de/newsticker/meldung/Kuenstliche-Intelligenz-in-der-Medizin-Wir-wollen-Aerzte-nicht-arbeitslos-machen-3824121.html> accessed 27 June 2018.

[94] DFK BitKom, 'Künstliche Intelligenz' (n 13) 61, 115.

[95] Martini, 'Algorithmen' (n 44) 1019.

tient. Human beings must not lose control of the decision-making itself, ie must not blindly trust an AI-decision. The value of AI-systems in healthcare lies in supporting the physicians' decision-making process, not completely giving up the decision-making authority.[96] We must not lose humanity and self-determination in such a sensitive field like healthcare. One important component will be to make automated decision-making transparent and understandable. Therefore, we need the right to explanation of a specific decision.

[96] DFK BitKom, 'Künstliche Intelligenz' (n 13) 56.