Gryz, Jarek; Rojszczak, Marcin

Article

Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Gryz, Jarek; Rojszczak, Marcin (2021) : Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 10, Iss. 2, pp. 1-24,
https://doi.org/10.14763/2021.2.1564

This Version is available at:
https://hdl.handle.net/10419/235967

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 10 **Issue** 2

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Black box algorithms and the rights of individuals: no easy solution to the "explainability" problem

Jarek Gryz *York University* jarek@cse.yorku.ca
Marcin Rojszczak *Warsaw University of Technology* marcin.rojszczak@pw.edu.pl

DOI: https://doi.org/10.14763/2021.2.1564

Published: 30 June 2021
Received: 18 December 2020 **Accepted:** 22 April 2021

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Gryz, J. & Rojszczak, M. (2021). Black box algorithms and the rights of individuals: no easy solution to the "explainability" problem. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1564

Keywords: Right to explanation, Explainable AI, Algorithmic transparency, Certification framework

Abstract: Over the last few years, the interpretability of classification models has been a very active area of research. Recently, the concept of interpretability was given a more specific legal context. In 2016, the EU adopted the General Data Protection Regulation (GDPR), containing the right to explanation for people subjected to automated decision-making (ADM). The regulation itself is very reticent about what such a right might imply. As a result, since the introduction of the GDPR there has been an ongoing discussion about not only the need to introduce such a right, but also about its scope and practical consequences in the digital world. While there is no doubt that the right to explanation may be very difficult to implement due to technical challenges, any difficulty in explaining how algorithms work cannot be considered a sufficient reason to completely abandon this legal safeguard. The aim of this article is twofold. First, to demonstrate that the interpretability of "black box" machine learning algorithms is a challenging technical problem for which no solutions have been found. Second, to demonstrate how the explanation task should instead be completed using well-known and well-trialled IT solutions, such as event logging or statistical analysis of the algorithm. Based on the evidence exposed in this paper, the authors find that the most effective solution would be to benchmark the automated decision-making algorithms using certification frameworks, thus balancing the need to ensure adequate protection of individuals' rights with the understandable expectations of AI technology providers to have their intellectual property rights protected.

1. Introduction

Recent advances in the development of machine learning (ML) algorithms, combined with the massive amount of data used to train them, has changed dramatically their utility and scope of applications. Software tools based on these algorithms are now routinely used in criminal justice systems, financial services, medicine, research and even in small business. Many decisions affecting important aspects of our lives are now made by algorithms rather than humans. Clearly, there are many advantages to this transformation. Human decisions are often biased and sometimes simply incorrect. Algorithms are also cheaper and easier to adjust to changing circumstances.

But algorithms have not proven a panacea. Despite promises to the contrary, there have been several instances of bias and discrimination discovered in algorithmic decision-making (Buiten, 2019, p. 42), particularly disturbing in the case of criminal justice (Huq, 2019; Richardson et al., 2019). Of course, once discovered, such bias can be removed and algorithms can be validated as non-discriminatory before they are deployed. But there is still widespread uneasiness—particularly among legal experts—about the use of these algorithms. Most of these algorithms are self-learning and their designers have little control over the models generated from the training data. In fact, computer scientists were formerly not very interested in studying these models because they were (and are) often extraordinarily complex (the reason they are often referred to as "black boxes"). The standard approach was that as long as an algorithm worked correctly, no one bothered to analyse how it worked [1].

This approach changed once the tools based on ML algorithms became ubiquitous and began directly affecting the lives of ordinary people (Pasquale, 2015). If the decision about how many years one will spend in prison is made by an algorithm, the convicted should have the right to know *how* this decision is made. [2] In other words, there is a clear need for the transparency and accountability of automatic decision-making (ADM) algorithms (Larsson & Heintz, 2020).

In recent years, many published papers have addressed the *interpretability* (variously defined) of models generated by ML algorithms. It has been argued that inter-

1. This is how Chris Anderson summarised this approach: "*Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.*" (Anderson, 2008, n.p.)
2. Advanced algorithms have been used in criminal justice systems, both in the United States and increasingly in Europe (Završnik, 2019).

pretability is not a monolithic notion. As a result, the subjectivity of each interpretation, due to different levels of human understanding, implies that there must be a multitude of dimensions that together constitute interpretability (Chakraborty et al., 2017). However, Zachary Lipton (2018) suggests that not only is the concept of interpretability muddled, it is also badly motivated. The approval of EU regulation 2016/679 (General Data Protection Regulation or GDPR) in 2016 prompted discussion of a related legal concept, the *right to explanation*. If this right is indeed mandated by GDPR (in effect since 2018), then software companies conducting business in Europe [3] are immediately liable if they are not able to satisfy this right.

The aim of this paper is to answer the question of whether and to what extent—given the specificity of ML systems—it is possible to provide information that would demonstrate algorithmic fairness, and as a result, compliance with the right to explanation. The first section analyses the concept of explanation within its legal as well as psychological context. We then demonstrate—using a case study of a music recommendation system—that the interpretability of "black box" algorithms is a challenging technical problem for which no solutions have yet been found. To that end, we show that models created by ML algorithms are inherently so complex that they cannot be "explained" in a meaningful way to an ordinary user of such systems. Instead, rather than looking "inside" an algorithm, we propose focussing on its statistical fairness and correctness. A promising way to achieve this goal may be to introduce event logging mechanisms and certification schemes, which are currently being used very successfully in the IT sector.

2. What is the right to explanation

One of the goals of the GDPR was to adapt EU regulations to modern methods of data processing, such as cloud computing or big data. [4] Hence, the EU legislature introduced a number of new provisions—including the widely discussed right to data portability (de Hert et al., 2018)—and expanded existing regulations (Hoofnagle et al., 2019), such as provisions on the right to information and automated decision-making.

3. It should be remembered that, due to the so-called territorial scope of application, the provisions of the GDPR should also be applied by entities having their headquarters in third countries (that is, outside the EEA) but directing their services to the market of at least one of the member states (de Hert & Czerniawski, 2016). The issue of the cross-border application of the GDPR is another practical problem in the enforcement of EU data protection legislation (Greze, 2019).

4. It is disputable to what extent this goal has been achieved. Tal Zarsky points out that "the GDPR fails to properly address the surge in Big Data practices. The GDPR's provisions are—to borrow a key term used throughout EU data protection regulation—incompatible with the data environment that the availability of Big Data generates" (Zarsky, 2017, p. 996).

According to the EU data protection model, every person has the right to know both the scope of data processed about them and the purpose of such processing. Furthermore, the data controller is required to provide them with this information "in a concise, transparent, intelligible and easily accessible form, using clear and plain language" (GDPR, 2016, Art. 12(1)).

In the EU legal system, the right to the protection of personal data—as well as the right to privacy—have been included in the catalogue of fundamental rights (CFR, 2012). Furthermore, it should be noted that, although both rights are closely related, they are, in fact, independent rights. This means, in particular, that—at least in the scope of EU law—data protection laws may be infringed even if privacy has not been affected in any way. Undoubtedly, one of the main goals of establishing dedicated data protection regulations is to guarantee the rights and freedoms of individuals in the digital era, and protect them from new types of threats arising from rapid technological development and the globalisation of modern IT services.

Article 22 of the GDPR is aligned with this goal; it introduces the right to not be subject to a decision made as a result of automated data processing that legally affects an individual or otherwise has a significant impact upon them. This regulation was also enshrined in Directive 95/46, the GDPR's predecessor, which was in place for over 20 years. However, since bulk algorithmic processing of personal data has developed rapidly only within the last two decades, the practical significance of this provision was insignificant. The situation has changed with the growth in profiling, including profiling for purposes other than advertising products and services (*Data Is Power*, 2017). It is worth noting that Article 22 of the GDPR does not explicitly provide for an individual's right to explanation of an automated decision. Instead, it sets out the general principle that an individual may object to automated decision-making (Malgieri & Comandé, 2017, p. 246).

In the case of automated decision-making, the EU legislature has extended the information obligation imposed on data controllers by introducing in Article 15(1)(h) of the GDPR the need to provide "meaningful information" on the logic involved in such decisions, taking into account the "significance and the envisaged consequences of such processing for the data subject". And it is this regulation that is the source of the term "right to explanation", though the phrase itself does not appear directly in the wording of the regulation. This interpretation is confirmed by Recital 71 of the GDPR, which states that processing based on automated decisions should always be subject to suitable safeguards, including the "right to obtain an explanation of the decision reached after such assessment and to challenge the decision".

Hence, the question arises at the outset as to whether the right to explanation is in fact a separate (*per se*) right of an individual or just an element of a broader right—the right to information. Some scholars have questioned the very existence of such a right (Wachter et al., 2017), while others have pointed out that, regardless of how the right to explanation is defined, it is not "illusory" (Selbst & Powles, 2017). Undoubtedly, the right to explanation serves a specific purpose—to enable an individual to challenge the correctness of a decision that has been made by an algorithm. Without understanding what criteria and factors the decision was based on, this entitlement can not be exercised in practice. Indeed, failure to provide a procedure to challenge the decision, including legal action, would deprive individuals of a key fundamental right—the right to a fair trial. It should be noted that, within the GDPR, only automated decisions that legally affect or otherwise significantly impact an individual are addressed. This is an important condition, the omission of which may lead to false conclusions about the legal scope of the right to explanation. However, Michael Veale and Lilian Edwards, referring to the Article 29 Working Party's position, advise a broad interpretation of this condition by showing that commonly used price comparison online services can also have "significant effects" on individuals (Veale & Edwards, 2018, p. 401).

The term "to obtain an explanation" used in the context of an automated decision may suggest that the obligation of a controller using automated decision-making is to explain how the algorithm reaches a specific result, which, according to Article 13(1) of the GDPR, should be presented in a transparent and intelligible form, using "clear and plain language". A significant part of the controversy surrounding the right to explanation relates precisely to the possibility of meeting this condition.

Before trying to identify the source of the difficulty, the term "explanation" in the context of decision-making needs to be clarified. Decision-making tools are based almost exclusively on classification algorithms. Classification algorithms are "trained" with data obtained from past decisions to create a model which is then used to arrive at future decisions. In this case the model requires an explanation, not the algorithm itself (in fact, different algorithms may be generated by very similar models).

When a user submits their information to a decision-making tool, an answer is generated—such as a number, a *No*, or a category such as "high risk". From the wording of Recital 71 (which states that the user has the right *to challenge the decision*) it is clear that the right to explanation is provided for cases where the answer given by the tool is different from what the user expected or hoped for. The

most straightforward question an individual may then ask is: "Why X?". When the user asks "Why X?", having expected a different answer ("Y"), they mean in fact to ask: "Why X rather than Y?". This type of question calls for a *contrastive explanation* (Miller et al., 2017). The answer that needs to be provided to the user must contain not only the explanation as to why the information provided by the user generated answer X, but also what information must change in order to generate answer Y (the one the user was expecting).

When people ask "Why X?", they are looking for the cause of X. Thus, if X is a negative decision for a loan application, an answer would need to specify what information in an application (the so-called "features" used as input in the model) caused X. It should also be remembered that the decision-making tool making a decision for a user is *replacing* a human that used to make such decisions. In fact, a person reporting a decision to the user may not clearly state that the decision is the verdict of an algorithm (judges in the US routinely use software-based risk assessment tools to help them in sentencing). The user may thus expect that the explanation provided uses the language of *social attribution* (Miller et al., 2017), that is, explains the behaviour of the algorithm using folk psychology.

3. A case study: Building a music recommendation system

As it was argued in the introduction, algorithm interpretability is a challenging task for their designers. Three barriers to the transparency of algorithms in general are usually distinguished: (1) intentional concealment whose objective is the protection of intellectual property; (2) lack of technical literacy on the part of users; (3) intrinsic opacity which arises from the nature of ML methods. A right to explanation is probably void when trade secrets are at stake (see Recital 63 of the GDPR; see Article 29 Working Party, 2017, p. 17), but the other two barriers still need to be addressed. In fact, these two barriers depend on each other. The complexity of ML methods positively correlate with the level of technical literacy required to comprehend them.

The most obvious solution to the second barrier would be implementing educational programmes aimed at transferring knowledge about the functioning of modern technologies. This could be achieved with stronger education programmes in computational thinking, and by providing independent experts to advise those affected by algorithmic decision-making (Lepri et al., 2018). The effectiveness of this solution, however, is questionable: even if it were possible to improve technical literacy education (which seems very unlikely given previous experience in this

area), that still leaves 80% of the population who completed their education many years ago.

As a solution to the last barrier, namely, the lack of transparency relating to the nature of ML methods, some sort of evidence gathering based on registering the key parameters of the algorithm should be sufficient (Wachter et al., 2017). Indeed, collecting this type of data would certainly help to understand how a system arrived at a specific decision. That said, it would still be completely unrealistic to expect a layperson to grasp these concepts.

Over the last few years much work has been done on "black box" model explanation. Some of this work (Adler et al., 2016; Baehrens et al., 2010; Lou et al., 2013; Montavon et al., 2018; Simonyan et al., 2014; Vidovic et al., 2015) has been aimed specifically at experts. The interpretability of a model is a key element of a robust validation procedure in applications such as medicine or self-driving cars. But there has also been some innovative work on model explanation alone (Datta et al., 2016; Fong & Vedaldi, 2017; Lakkaraju et al., 2019; Ribeiro et al., 2016, 2018; Shrikumar et al., 2016; Tamagnini et al., 2017; Yosinski et al., 2015; Zintgraf et al., 2017). Most of these papers are addressed to experts, with the aim of providing insights into the models they create or use. In fact, only in the last three papers mentioned above were explanations tested on people, and even then a certain level of sophistication was expected on their part (from the ability to interpret a graph or bar chart to completing a postgraduate course on ML). Most importantly, though, all of these works provide explanations of certain *aspects* of a model (for example, showing what features or attributes most influence the decision of an algorithm). None of them attempt to explain fully the two contrasting paths ("why X rather than Y") in a model that lead to distinct classification results (which, as stated above, is necessary for a contrastive explanation).

Indeed, explaining the black box model of an ADM algorithm is much harder than is normally assumed. To illustrate this case better, we describe in this section recent work we were involved with (Shahbazi et al., 2018) on designing a song recommendation system for KKBOX, Asia's leading music streaming service provider.

KKBOX had provided a training data set that consisted of information from listening sessions for each unique user-song pair within a specific timeframe. This information available to the algorithm includes information about the users, such as identification number, age, gender, etc., and about songs, such as length, genre, singer, etc. The training and the test data were selected from users' listening history in a given time period and had around 7 and 2.5 million unique user-song pairs

respectively.

The quality of a recommendation system's predictions relies on two principal factors: predictive features available from past data (for example, what songs the user has listened to the most) and an effective learning algorithm. Very often, these features are only implicit in the training data and the algorithm is not able to extract them by itself. Feature engineering is an approach that exploits the domain knowledge of an expert to extract from the data set features that should generalise well to the unseen data in the test set. The quality and quantity of the features have a direct impact on the overall quality of the model. In this case, certain statistical features were created (or extracted, because they were not explicitly present in data), including the number of sessions per user, the number of songs per session and the length of time a user had been registered with KKBOX.

As a result, the number of features available to the algorithm was increased by a factor of about 10, to 185. And this is the key point: some of these derived features turned out to be extremely important in determining a user's taste in songs and, as a result, the recommendation that was provided. But it should be emphasised that none of these features were explicitly present in the original data. The paradox is that if someone asked for an explanation of how the model worked, the answer would have to be based on features *not* present in the source data.

But this is only part of the story. The solution provided did not use a *single* algorithm to make a prediction. In total, five different algorithms were used, all of them very complex. Thus, here is another key point: the final model was the weighted average of all five models' predictions. Again, it should be stressed that the result was not the outcome of just one algorithm. Figure 1 shows the complexity of one of these algorithms in the form of a simplified neural net structure. [5]

5. Each of these steps has not been explained in detail as the key point is simply to present the complexity of the entire prediction process, not its technical aspects.

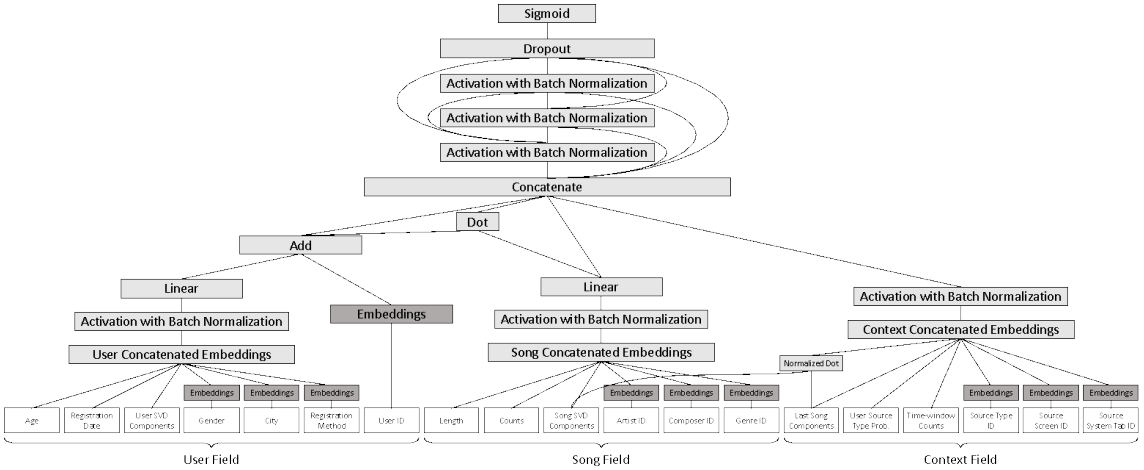

FIGURE 1: Structure of one of the algorithms used in the recommendation system (Shahbazi et al., 2018)

The model that was generated by these algorithms was also extremely large and complex. Since gradient boosting decision tree algorithms were used, the resulting model was a forest of such decision trees. [6] The forest contained over 1,000 trees, each with 10-20 children at each node and at least 16 nodes deep. [7]

The question arises as to how a user can understand this model. One can begin by assuming that a user wants an explanation for why song X was recommended rather than song Y. There will be multiple trees with the X recommendation as well as the Y recommendation. But which one offers the right choice? These multiple trees cannot be generalised as this has already been done by the algorithm (one of the most difficult aspects of algorithms based on decision trees is their optimisation, that is, generating the simplest, most general trees). Indeed, an ordinary user would not be able to comprehend the model, let alone understand an explanation that uses vocabulary entirely foreign to them. It is up to the experts to verify the explanation and convey this verification to the user.

The ADM models are often even more complex than the system described above. Machine learning is heuristics-driven and no one expects rigorous mathematical proofs of the correctness of its algorithms. What often happens is that, if a model generated by an algorithm does not correctly classify the test data, a designer will place another algorithmic layer on top of it in the hope that it improves the re-

6. Nodes in a decision tree store conditions that have to be satisfied (for example, she must be under 15 years of age) if a user is to be recommended a particular song.
7. It took almost 128GB of RAM to derive the gradient boosting decision tree model and around 28 hours on 4 Tesla T4 GPUs to create the deep neural network model.

sults. Sometimes it does but at this point no one would be able to explain why this had happened. As Ali Rahimi put it in a recent keynote talk at the Conference on Neural Information Processing Systems (Rahimi, 2017, n.p.): "Machine learning has become alchemy (…) many designers of neural nets use technology they do not really understand". If the people who design these algorithms do not understand them, how can anyone else?

4. Who needs the explanations anyway?

The juxtaposition of legal requirements arising from the GDPR with the specificity of ML systems has led to serious doubt about the actual usefulness of the right to explanation of an automated decision. Proponents of the view that the right to explanation is useless in the world of machine learning systems highlight two important arguments: one of a technological nature and the other of a social nature. First of all, as stated above, the way ML systems work makes it difficult (or even impossible) to present the criteria used by an algorithm when resolving a given case. It should be remembered that the decisions made by ML systems largely depend on the data used in the system learning process (this is related to the so-called incremental effect). [8] This conclusion is based not only on the presumption that understanding algorithms is too difficult for people, but also on the fact that, in general, the way algorithms operate and process information is qualitatively different from how humans operate and process information and, as such, the term "interpretability" has a different meaning both for people and ML algorithms (Krishnan, 2019). However, even if the technological limitation is overcome, another problem becomes apparent: the average individual's lack of knowledge and expertise in analysing and evaluating the very complex results of operations carried out by advanced ML algorithms, where highly specialised knowledge is needed.

The latter issue will be analysed first. It can be reduced to the following argument: *It is not necessary to explain the decisions made by the algorithms because no one will understand the explanation in the first place*. If this were true, the same reasoning could be applied to the problem of analysing flaws and defects related to the operation of other advanced systems and products, such as cars and airplanes. Most users do not understand how a CPU works, but they are not denied the right to de-

8. The incremental effect consists of changing the operation of the algorithm as a result of providing new information to the database. The algorithm "learns" on the basis of the new information, which may lead to a different interpretation of the information processed previously. Hence, the result of the algorithm is variable over time, which means that by providing the same data for analysis, different outcomes can be obtained. This leads to the conclusion that, in the case of ML algorithms, attempting to confirm their correct operation by processing the same data set at another time is not a good strategy.

termine whether it was a processor failure that caused a plane to crash. Technology is becoming more and more complex every year, and this is true not only of the IT world. Most people do not understand the medical therapies they undergo, economic processes that affect their financial position or legislation—even though they are obliged to abide by it. At the same time, if an individual considers that they have suffered harm, or that their rights have been undermined, they can take their case to court. One does not have to be a professor of medicine to claim compensation for medical malpractice. [9] The scope or existence of this right should not be contingent upon whether the wrong diagnosis was made by a medical practitioner or by an algorithm. If the court decides that expert knowledge is needed to resolve a given case, it will appoint expert witnesses to assess the evidence gathered in that case. In this way, expert witnesses can help determine the causes of a plane crash, whether medical malpractice took place, or who has liability for a leaking roof in a house. Experts familiar with modern decision-making systems should be able to analyse the results of an algorithm's operation in the same way. [10] However, for this to be possible, individuals affected by such a decision must have the right to know how this decision was reached. Depriving them of this right would effectively condone the practice of unknown decision-makers making non-transparent decisions according to unknown criteria, with no real possibility of challenging such decisions. This is a Kafkaesque world, incompatible with the principles of a democratic society.

5. Possible (and feasible) solutions

Assuming a general consensus that an individual should be able to challenge decisions taken automatically, the next step that needs to be addressed is to overcome the technical difficulty in determining (reconstructing) the criteria that were taken into account by the algorithm while formulating its decision. This problem should not be underestimated. As illustrated in Section 3, a relatively simple recommendation system used by a music provider demonstrates that in the era of big data systems, even seemingly straightforward decisions ("*which song to recommend to a user*") are made with the use of very advanced algorithms. Society expects that IT systems will work not only faster than people, but also more efficiently and effectively, which means that algorithms will be able to solve complex problems with a

9. It should be remembered that nowadays medicine is one of the main areas of application for ML algorithms (Hoeren & Niehoff, 2018).

10. *Cf* the examples discussed by Jenna Burrell, which she uses to "illustrate how the workings of machine learning algorithms can escape full understanding and interpretation by humans, even for those with specialized training, even for computer scientists" (Burrell, 2016, p. 10).

speed unattainable for humans, and that they will also be able to solve problems that people could not otherwise solve at all (Hecht, 2018). Algorithm predictions are made in all applications of ML systems, including those extremely critical for individuals, such as medical diagnostics (Hoeren & Niehoff, 2018). However, due to the almost complete opacity of algorithm functioning, any attempt to trace their mode of operation, even by an expert in the field, if not actually impossible, would be affected by such a large margin of error as to make any results wholly unreliable (Burrell, 2016). In order to understand the correctness of a decision, an expert or even a group of experts, would have to not only learn the logic of the algorithm but also trace previous decisions and familiarise themselves with the system's learning (training) process. Due to the increasing complexity of this type of algorithm, the scale of this problem will only escalate.

Providing an explanation that is understandable to humans also requires assessing the quality of the data on which an algorithm is based. Classification algorithms need data to learn how to make predictions. This training set must be representative of that data and sufficiently large. For example, the data set for the KKBOX recommendation system described in Section 3 contained information on 30,000 users, 360,000 songs and 7 million user-song pairs. One of the main sources of AI success has been the emergence of 'big data', that is, freely and automatically collected data widely available for anyone to use. However, it is important to note that the amount of data alone is not sufficient to generate correct predictions; the data must also be representative. In ADMs the problem may be further compounded by uncritical analysis, leading to discriminatory conclusions (Barocas & Selbst, 2016).

The data used by ADMs must therefore be validated to ensure lack of bias. Obviously, this is not an easy task. First, the data sets used by ADM systems are huge and cannot be analysed "manually". To automate this process, the type of bias that might impact further processing should be defined in advance. Second, most of the data used by ADMs stems from past decisions made by humans, which could conceivably be biased along racial or gender lines. Therefore, when considering possible technical implementations of the right to explain in the context of ADM, the problem of ensuring adequate quality of data should also be addressed. In short, it is necessary not only to analyse the mechanisms used for confirming the correctness of an algorithm itself, but also the existence of safeguards that ensure the processed data is trustworthy.

There are at least two possible solutions to this problem. The first would require mandatory registration of the key parameters of those ADM systems whose deci-

sions have legal ramifications for individuals (as in the case of Article 21 of the GDPR). The second way to validate the operation of an algorithm is not so much an attempt to trace the correctness of its decisions as a formal evaluation of the entire system through certification measures. The following sections will discuss both proposals, together with an analysis of their main advantages and limitations.

5.1. An event logging subsystem

A proven solution, used by IT system designers in cases where it was necessary to trace (reconstruct) the operation of an algorithm at a later stage, is the recording of significant processing parameters. A typical example of such a mechanism are flight recorders, the key elements used to determine the course of flight events. This proposal therefore aims to introduce an obligation to record (log) the reasons for decisions made by an ML algorithm. Proponents of such a solution highlight the ability to trace the correct operation of the system and thus the accuracy of the conclusions reached—what Margot Kaminski describes as "qualified transparency": to provide individuals, experts and regulators with different, but appropriate, sets of information related to algorithmic decision-making (Kaminski, 2019).

The recording of relevant parameters is relatively simple to implement, does not increase the costs of deploying and maintaining the system, and does not require time-consuming validation procedures. These are important benefits because, when considering any proposals related to fulfilling regulatory requirements, one should not lose sight of their economic consequences. ML systems are mostly developed for global application. The introduction of regulations whose implementation would require significant costs to be borne by technology providers could lead to a distortion of market competition or result in providers' relocation to jurisdictions where such regulations have not been implemented.

In addition to being straightforward to implement, the logging of system parameters can also be easily secured cryptographically to ensure the consistency and integrity of recorded data. Taking into account the type of ML system or sensitivity of data processed, logs can be maintained by a specific service provider or trusted third party—avoiding the risk of the data being changed without authorisation Moreover, there is no obstacle to such data being stored in systems supervised by public entities; in this way, the relevant parameters of, for example, a machine-based credit scoring system could be securely stored under the oversight of a financial market supervisor. This, in turn, opens up the possibility of introducing sector-specific requirements that would define a minimum set of parameters to be recorded by automatic decision-making systems and used for the provision of ser-

vices in regulated markets. Under this approach, a person challenging the correctness of a decision taken or wishing to exercise their right to explanation of an automatic decision (Article 22 of the GDPR) would have access to the set of key parameters that influenced the final decision. In turn, the supervisory authority could have access to a wider (and more detailed) set of parameters with which it could analyse not only individual cases but also the regularity and legality of the operation of the whole system.

The solution outlined above does have its weaknesses. First of all, it cannot be applied to all types of machine learning algorithms—in particular, deep neural networks with weights attached to features and complex interactions that are not directly interpretable, and therefore no user-interpretable arguments that can be recorded.

ML systems are also not 'static'—with new data, the prediction model generated by an algorithm will change. As a result, the inference process will be modified (e.g. new parameters will be included or pre-existing parameters omitted) and event logging mechanisms will change as well. In traditional IT solutions, it is the main user of the system who determines the set of data to be recorded and also indicates how often such recording should be done. Both the scope of data and the frequency of ML recording are criteria which cannot be defined in advance. Practically speaking, it is the system itself (or one of its components) that should be designed to determine what parameters are to be recorded and when. However, this goes against the idea behind this safeguard—to ensure transparency. Since it is not the developer who would establish strict and unchangeable criteria for recording key parameters, but the system itself, this mechanism could also be prone to error or external manipulation. As a result, there would need to be a formal evaluation of the recording process itself. In other words, the attempt to solve the problem of the transparency of an ML system would be replaced by the problem of ensuring the transparency of the event logging subsystem.

Another limitation of this solution is the context of analysis, which is difficult to take into account. It should be remembered that the operation of an algorithm depends not only on the input data and internal procedures for processing (the result of which is also easy to save), but also on previous analyses—that is, on the whole tree of decisions made earlier. Understanding the current result of an algorithm may therefore require the review of a huge knowledge base describing previous decisions made by the system. Without this information, simply saving the current parameters used in the inference might not allow one to reconstruct (and thus verify the correctness and fairness of) the inference performed. The more an algo-

rithm is based on machine learning mechanisms, the more this problem will make difficult the use of logging as a way of ensuring system transparency.

A third limitation that needs discussing is the unobvious relationship between the stored parameters and the internal logic of an algorithm. Even assuming that the two previously mentioned obstacles can be overcome, and that the recording of key parameters allows the full and precise reproduction of the initial state and results of subsequent processing steps, the problem of access to the internal logic of an algorithm will subsequently become apparent. ML systems, like other highly specialised technologies, are subject to intellectual property protection (Gervais, 2020). The effectiveness of the protection of various AI technology components is a significant problem affecting the growth of this market. Without access to the source code—and thus to the logic of an AI algorithm—even detailed parameters of its operation will not be sufficient to fully understand the decision-making process whose correctness is to be assessed.

Another issue to be clarified is the adequacy of this measure in achieving its intended purpose. In fact, advocates of the transparency of processing expect the reliability (credibility) of algorithms' operation to be ensured. It seems, however, that ensuring the transparency of the system will not always be a sufficient guarantee of processing reliability—and thus the protection of an individual's rights. Ensuring that processing is fair must include not only confirmation of the correctness of the processing carried out but also its compliance with legal or ethical standards. After taking into account these additional limitations, it may turn out that a properly functioning IT system, which identifies objectively correct relationships between data, cannot be considered trustworthy. It will not be possible to reveal this limitation solely by recording the processing parameters. These parameters alone will not reveal a defect relating to the external data on which an algorithm is based.

5.2. Certification frameworks

A second way to validate the operation of an algorithm is not so much an attempt to trace the correctness of its decisions, then a formal evaluation of the entire system through certification. It proposes the creation of a national (or international) certification framework for machine learning systems. The purpose of such a framework would not only be to ensure that systems used to make automated decisions were designed, built and tested in compliance with applicable norms and standards, but also to make sure that their mode of operation (the reliability of decisions made) was confirmed statistically.

In the IT industry, certification mechanisms have been used for years to confirm the authenticity and integrity of software systems (Heck et al., 2010). The use of an external certification mechanism (independent of the provider or user) in relation to machine learning systems could also help to eliminate the risk of unauthorised interference in the way a system works. Furthermore, certification would not have to be mandatory—it could be an optional measure. To encourage ML system providers to participate in this framework, the legislature could introduce a number of legal presumptions based on the premise that decisions made by a certified system are correct. As with any legal presumption, a party challenging such a decision could contest it in court, but they would be required to prove the malfunction of the system. Certification would therefore be a mechanism that obviates the necessity to later prove the correctness and fairness of a system in litigation.

The proposal to introduce certification of advanced IT systems is not a new one and has already been defined, for instance, in relation to artificial intelligence (AI) systems. Matthew Scherer (2016), suggested regulating the AI market with a supervisory body that would issue certifications for AI systems (including tests of new versions of software agents). According to his proposal, certification was not a prerequisite for putting a system into operation but rather a manifestation of *soft law* regulation. This would provide an incentive for developers by limiting the liability for damage caused by their systems (Scherer, 2016). A similar idea was mooted 20 years earlier by Curtis Karnow. The model he proposed was simpler and primarily involved the creation of the *Turing Registry* (a hypothetical list of "safe" AI agents), without a reference to any regulatory aspects (Karnow, 1996).

It is worth noting that the implementation of a certification framework for systems making automated decisions is a solution that can be reconciled with the current wording of GDPR provisions. An element of every formal IT system certification framework is an assessment of whether the documentation provided is complete and up to date. It can be expected that in the case of ML systems, such documentation would contain not only a technical description of the environment and the algorithms used, but also a high-level description of the system's operating principles—prepared in a simple and readable manner, compliant in this respect with Article 15 of the GDPR.

It appears, therefore, that the introduction of a certification framework may be helpful in solving both of the problems discussed above. On the one hand, this solution would take into account the specificity of ML systems and would be technically feasible; on the other, it would not require people who want to challenge automated decisions to have specialised knowledge in the field of data analysis or

the structure of expert systems.

However, the proposal to use certification frameworks also requires the resolution of several important problems. Firstly, it should be remembered that different certification mechanisms are used in the IT industry. In general, they can be divided into those confirming the correctness of software development and maintenance processes (process certification) and those intended to confirm the authenticity and integrity of software (code certification) (Eloff & von Solms, 2000). In both areas, different norms and standards are used.

Code certification makes it possible to ensure that no third party has interfered with and changed the structure of the computer software. However, such certification only applies to software supplied (or implemented) by the manufacturer (developer), and therefore does not confirm lack of interference with the memory structure of the ML system being run. In particular, it does not in any way refer to the possibility of poisoning the ML logic by deliberate manipulation or feeding the system with badly prepared data. Although system certification mechanisms have been used in the IT sector for several decades, they have so far been used mainly to validate systems that process sensitive data, e.g. in the area of state security (Lipner, 2015). This is due to the simple fact that formal certification of an IT system is a very time-consuming and costly process (Kaluvuri et al., 2014). The wide application of the existing certification framework, such as the Common Criteria (ISO/IEC, 2009), is therefore not enough to fully reflect the needs of the ML market, and it also seems problematic for commercial reasons (see generally, Mellado et al., 2007). It is difficult to imagine that European technology providers would conduct formal certification that might delay their product launch onto the market, whereas the activities of entities operating in other jurisdictions would not be limited in this way.

With regard to process certification in the IT industry, for years the reference frameworks have been the ISO/IEC 20000 and ISO/IEC 27001 family of standards (Siponen & Willison, 2009). Management systems built on their basis may be subject to formal certification. However, it should be remembered that in this scenario certification would ensure that the development, implementation and maintenance of IT systems were carried out with best practice in mind, and in a way that minimised identified risks. Moreover, management systems are part of *soft law* regulation, so they are mainly the source of internal requirements in the compliance area of the service provider and do not lay down legally binding obligations towards the system users. Processes' certification can also be used to establish a secure supply chain, in which many actors are *de facto* responsible for the proper op-

eration of an ADM system. In this case, it would be possible to introduce standards dedicated to particular categories of entities, e.g. data brokers, companies responsible for data cleaning and quality assurance processes or those involved in the ADM training process. These standards could be subject to a formal evaluation of conformity by an independent external body in a similar way to current certification of management systems.

While certification is a good way to regulate the introduction and operation of ADM systems, there are currently no certification schemes that can be applied directly to this end. What is more, there are not even any legal regulations—at either EU or member state level—that could form the basis for introducing such certification schemes. Even Regulation 2019/881, which creates a framework for certification in the area of cybersecurity, cannot be regarded as such. The main application of the regulation is to improve the security of products used by critical infrastructure operators and digital service providers (Rojszczak, 2020). The main area of application of ML systems, in turn, is the mass consumer market. Hence, it seems that before it is possible to address in detail a future certification framework for ADM systems, it will be necessary to discuss the establishment of new EU regulations that could form the basis of such programmes.

Reuben Binns (2018) aptly notes that current approaches to fair machine learning are typically focused on interventions at the data preparation, model-learning or post-processing stages. Although certification seems to be a promising solution to the problem of confirming the correct operation of ADM algorithms, it will not overcome the significant limitation *strictly* related to the very nature of statistical analysis. As noted earlier, the right to an explanation is seen not only as a means of confirming the correctness of the decision but also a means of establishing the reasons for not taking the decision that the applicant had expected (the "*why X and not Y?*" problem presented earlier). As a result, even if a specific algorithm generates statistically correct results, which are confirmed in the certification procedure, its operation can still be questioned because an individual will be deprived of the possibility of ascertaining what circumstances determined the unexpected, or unwelcome, outcome.

6. Conclusions

Black box algorithms make decisions that affect human lives. This trend is not expected to change in the coming years. Automatic decisions will be made not only on an ever-increasing scale, but also with ever-increasing intensity—as a result of which there will also be increasing pressure on public opinion to develop effective

control mechanisms, including those which make it possible to question the decision made in individual cases.

Numerous researchers have criticised the very concept of a right to explanation, pointing out the lack of precision of the EU legislature (Wachter et al., 2017) and questioning the usefulness of this right in practice (Edwards & Veale, 2017). Due to different definitions of and approaches to the "explainability" problem of ML systems, Cynthia Rudin (2019, p. 206) has stated that "the field of interpretability/explainability/comprehensibility/transparency in ML has strayed away from the needs of real problems."

Today, the right to explanation of an automated decision may be perceived as one of the less important elements of the GDPR, with limited practical significance. However, this perception will change soon. ML systems are entering new areas of the economy as well as public administration. Hence, the wording and limits of applicability of the law laid down in the GDPR will undoubtedly be subject to recurrent interpretation, including interpretation by the Court of Justice of the European Union.

This would therefore seem an opportune moment to begin discussing the need for a comprehensive regulation on how ML systems are developed, implemented and supervised. Drawing on the experience of the IT sector, it seems most appropriate to introduce a regulatory model in which various types of certification mechanisms will play a leading role. The basis for such a model may be a certification scheme for ML systems—allowing for different certification schemes for systems operating in different markets. It will certainly be necessary to distinguish a specific category of systems whose decisions may affect fundamental rights and freedoms. Future legislation should also promote the use of *soft law* measures, such as certification based on international standards or codes of conduct, to support the development of industry standards and self-regulation mechanisms. An example of such soft law is the ISO/IEC CD 23053 (2020), a draft international standard that is intended to establish a framework for artificial intelligence systems using machine learning. Regardless of the certification, in the case of less advanced ML systems, it may be sufficient to use standardised (e.g. resulting from recommendations issued by competent supervisory authorities) procedures for recording key systems parameters. This proposal may additionally be combined with the establishment of a dedicated supervisory authority, competent to moderate the development of an AI market and—by introducing various regulatory mechanisms, including certification—ensuring their safe use (Tutt, 2016).

It should not be expected therefore that a single, universally-accepted certification scheme for ADM systems will be developed. It is also unlikely that such a uniform standard will be developed within the EU in the near future. The reason for this is not only the lack of consensus between member states on the need to establish EU regulation in this area but also the different digital maturity of individual national markets. Hence, it seems more probable that a set of different legal safeguards which can be applied in particular EU countries will be developed in order to ensure that the dynamic development of technology—including the spread of ADM—does not adversely affect the area of fundamental rights. This trend is already being observed today (Malgieri, 2019), and the problem of implementing the right to explanation of decisions taken automatically is one of the main areas of legislative activity.

References

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016, December). Auditing black box Models for Indirect Influence. *2016 IEEE 16th International Conference on Data Mining (ICDM*. https://doi.org/10.1109/icdm.2016.0011

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Article 29 Working Party. (2017). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01)*.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to Explain Individual Classification Decisions. *Journal of Machine Learning Research*, *11*, 1803–1831. https://www.jmlr.org/papers/volume11/baehrens10a/baehrens10a.pdf

Barcoas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 149–159. http://proceedings.mlr.press/v81/binns18a.html

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI* (pp. 1–6). https://doi.org/10.1109/UIC-ATC.2017.8397411

Charter of Fundamental Rights of the European Union, (2012).

Data is power: Towards additional guidance on profiling and automated decision-making in GDPR.

(2017). [Report]. Privacy International. https://privacyinternational.org/report/1718/data-power-profiling-and-automated-decision-making-gdpr

Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *2016 IEEE Symposium on Security and Privacy (SP)*. https://doi.org/10.1109/sp.2016.42

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*, *16*(1), 18–84. https://dltr.law.duke.edu/2017/12/04/slave-to-the-algorithm-why-a-right-to-an-explanation-is-probably-not-the-remedy-you-are-looking-for/

Eloff, M. M., & Solms, S. H. (2000). Information Security Management: An Approach to Combine Process Certification And Product Evaluation. *Computers & Security*, *19*(8), 698–709. https://doi.org/10.1016/S0167-4048(00)08019-6

Fong, R. C., & Vedaldi, A. (2017, October). Interpretable Explanations of Black Boxes by Meaningful Perturbation. *2017 IEEE International Conference on Computer Vision (ICCV*. https://doi.org/10.1109/iccv.2017.371

Gervais, D. (2020). Is Intellectual Property Law Ready for Artificial Intelligence? *GRUR International*, *69*(2), 117–118. https://doi.org/10.1093/grurint/ikz025

Greze, B. (2019). *The extra-territorial enforcement of the GDPR: A genuine issue and the quest for alternatives*. International Data Privacy Law. https://doi.org/10.1093/idpl/ipz003

Hecht, J. (2018). Managing expectations of artificial intelligence. *Nature*, *563*(7733), 141–143. https://doi.org/10.1038/d41586-018-07504-9

Heck, P., Klabbers, M., & Eekelen, M. (2010). A software product certification model. *Software Quality Journal*, *18*(1), 37–55. https://doi.org/10.1007/s11219-009-9080-0

Hert, P., & Czerniawski, M. (2016). Expanding the European data protection scope beyond territory: Article 3 of the General Data Protection Regulation in its wider context. *International Data Privacy Law*, *6*(3), 230–243. https://doi.org/10.1093/idpl/ipw008

Hert, P., Papakonstantinou, V., Malgieri, G., Beslay, L., & Sanchez, I. (2018). The right to data portability in the GDPR: Towards user-centric interoperability of digital services. *Computer Law & Security Review*, *34*(2), 193–203. https://doi.org/10.1016/j.clsr.2017.10.003

Hoeren, T., & Niehoff, M. (2018). Artificial Intelligence in Medical Diagnoses and the Right to Explanation. *European Data Protection Law Review*, *4*(3), 308–319. https://doi.org/10.21552/edpl/2018/3/9

Hoofnagle, C. J., Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: What it is and what it means. *Information & Communications Technology Law*, *28*(1), 65–98. https://doi.org/10.1080/13600834.2019.1573501

Huq, A. Z. (2019). Racial Equity in Algorithmic Criminal Justice. *Duke Law Journal*, *68*(6), 1043–1134. https://scholarship.law.duke.edu/dlj/vol68/iss6/1

ISO/IEC. (2009). *ISO/IEC 15408-1:2009, Information technology—Security techniques—Evaluation criteria for IT security—Part 1: Introduction and general model.*

ISO/IEC, C. D. (2020). *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).*

Kaluvuri, S. P., Bezzi, M., & Roudier, Y. (2014). A Quantitative Analysis of Common Criteria Certification Practice. In C. Eckert, S. K. Katsikas, & G. Pernul (Eds.), *Trust, Privacy, and Security in Digital Business* (Vol. 8647, pp. 132–143). Springer International Publishing. https://doi.org/10.1007/978-3-319-09770-1_12

Kaminski, M. E. (2019). The Right to Explanation, Explained. *Berkeley Technology Law Journal*, *34*(1), 188–218. https://doi.org/10.15779/Z38TD9N83H

Karnow, C. E. A. (1996). Liability For Distributed Artificial Intelligences. *Berkeley Technology Law Journal*, *11*(1), 147–204. https://btlj.org/data/articles2015/vol11/11_1/11-berkeley-tech-l-j-0147-0204.pdf

Krishnan, M. (2019). *Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning*. Philosophy & Technology. https://doi.org/10.1007/s13347-019-00372-9

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019, January). Faithful and Customizable Explanations of Black Box Models. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314229

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1469

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*, 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lipner, S. B. (2015). The Birth and Death of the Orange Book. *IEEE Annals of the History of Computing*, *37*(2), 19–31. https://doi.org/10.1109/MAHC.2015.27

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD \textquotesingle13*. https://doi.org/10.1145/2487575.2487579

Malgieri, G. (2019). Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review*, *35*(5), 105327. https://doi.org/10.1016/j.clsr.2019.05.002

Malgieri, G., & Comandé, G. (2017). Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019

Mellado, D., Fernández-Medina, E., & Piattini, M. (2007). A common criteria based security requirements engineering process for the development of secure information systems. *Computer Standards & Interfaces*, *29*(2), 244–253. https://doi.org/10.1016/j.csi.2006.04.002

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *ArXiv*. http://arxiv.org/abs/1712.00547

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*.

Harvard University Press.

Rahimi, A. (2017). *NIPS 2017 Test-of-Time Award presentation*. Conference on Neural Information Processing Systems. https://www.youtube.com/watch?v=ORHFOnaEzPc

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.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD*. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). *Anchors: High Precision Model-Agnostic Explanations*. Thirty-Second AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review*, *94*, 15–55. https://www.nyulawreview.org/wp-content/uploads/2019/04/NYULawReview-94-Richardson_etal-FIN.pdf

Rojszczak, M. (2020). The Evolution of EU Cybersecurity Model: Current State and Future Prospects. In B. J. Pachuca-Smulska, E. Rutkowska-Tomaszewska, & E. Bani (Eds.), *Public and private law and the challenges of new technologies and digital markets* (Vol. 1, pp. 295–312). C. H. Beck.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353–400. http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–243. https://doi.org/10.1093/idpl/ipx022

Shahbazi, N., Chahhou, M., & Gryz, J. (2018). Truncated SVD-based Feature Engineering for Music Recommendation. *WSDM Cup 2018 Workshop, Los Angeles*. WSDM Cup 2018 Workshop, Los Angeles. https://wsdm-cup-2018.kkbox.events/pdf/2_WSDM-KKBOX_Nima_Shahbazi.pdf

Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *ArXiv*. http://arxiv.org/abs/1605.01713

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014, April 19). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations, Workshop Track Proceedings*. http://arxiv.org/abs/1312.6034

Siponen, M., & Willison, R. (2009). Information security management standards: Problems and solutions—ScienceDirect. *Information & Management*, *46*(5), 267–270. https://doi.org/10.1016/j.im.2008.12.007

Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017). Interpreting black box Classifiers Using Instance-Level Visual Explanations. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, 1–6. https://doi.org/10.1145/3077257.3077260

Tutt, A. (2016). An FDA for Algorithms. *Administrative Law Review*, *69*(1), 83–123. https://www.jstor.org/stable/44648608

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, *34*(2), 398–404. https://doi.org/10.1016/j.clsr.2017.12.002

Vidovic, M. M.-C., Görnitz, N., Müller, K.-R., Rätsch, G., & Kloft, M. (2015). Opening the Black Box: Revealing Interpretable Sequence Motifs in Kernel-Based Learning Algorithms. In *Machine Learning and Knowledge Discovery in Databases* (pp. 137–153). Springer International Publishing. https://doi.org/10.1007/978-3-319-23525-7_9

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. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Yosinski, J., Clune, J., Nguyen, A. M., Fuchs, T. J., & Lipson, H. (2015). *Understanding Neural Networks Through Deep Visualization*. http://arxiv.org/abs/1506.06579

Zarsky, T. (2017). Incompatible: The GDPR in the Age of Big Data. *Seton Hall Law Review*, *47*(4), 995–1020.

Završnik, A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*. https://doi.org/10.1177/1477370819876762

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *ArXiv*. http://arxiv.org/abs/1702.04595

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

centre internet et société

R&I IN3 Internet interdisciplinary Institute

Universitat Oberta de Catalunya