Futility of a Right to Explanation

Jarek Gryz
York University
Toronto
jarek@cse.yorku.ca

Nima Shahbazi
Mindle AI
nima@mindle.ai

ABSTRACT

In the last few years, 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 2018 EU introduced General Data Protection Regulation with a Right to Explanation for people subjected to automated decision making. The Regulation itself is very brief on what such a right might imply. In this paper, we attempt to explain what the Right to Explanation may involve. We then argue that this right would be very difficult to implement due to technical challenges. We also maintain that the Right to Explanation may not be needed and sometimes may even be harmful. We propose instead an external evaluation of classification models with respect to their correctness and fairness.

KEYWORDS

right to explanation, explainable AI, algorithmic transparency

1 INTRODUCTION

Recent advances in development of machine learning algorithms combined with massive amount of data to train them changed dramatically their utility and scope of applications. Software tools based on these algorithms are now routinely used in criminal justice system, 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 of this transformation. Human decisions are often biased and sometimes simply incorrect. Algorithms are also cheaper and easier to adjust to changing circumstances.

Yet there is a price to pay for these benefits. Despite promises to the contrary, there have been several cases of bias and discrimination discovered in algorithmic decision-making. Of course, once discovered, these biases can be removed and algorithms can be validated to be 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 not really interested in studying the models because they are often extraordinarily complex (hence they are often referred to as black boxes). The standard approach was that as long as an algorithm worked correctly, nobody bothered to analyze *how* it worked.

This approach has changed once the tools based on machine learning algorithms became ubiquitous and began directly affecting lives of ordinary people. If the decision on how many years you are going to spend in prison is made by an algorithm you have the right to know *how* this decision was made. In other words, we need transparency and accountability of the decision-making algorithms.

In recent years, multiple papers have been published to address *interpretability* (variously defined) of models generated by machine learning algorithms. However, a recent publication [8] suggests not only that the concept of interpretability is muddled but also badly motivated. The approval of General Data Protection Regulation (GDPR) 2016 prompted a discussion [1] on a related legal concept, *a right to explanation*. If this right is indeed mandated by GDPR (which has been in effect since 2018) then software companies conducting business in Europe are immediately liable if they are not able to satisfy that right. A discussion on what it would take to comply with this new requirement is thus already overdue.

In this paper, we discuss that very concept. Our conclusion is mostly negative; we do not believe that right to explanation can be successfully implemented or that it is useful. In Section 2, we set the stage for the discussion by defining precisely the context. In Section 3 we review recent work on model explanation and show that it has little relevance for implementing the right to explanation. Section 4 presents a case study of a recommendation system we have developed recently. We show that the model generated by algorithms of that system would be very difficult – if at all possible – to explain to an ordinary user. Thus, in Section 5, we contend that we do not need a right to explanation in the first place and show that in fact it can be harmful. We conclude the paper in Section 6.

We should also point out that most of the diagnoses and opinions expressed in this paper apply as much to a wider concept of *interpretability* as they do to *a right to explanation*.

2 WHAT IS A RIGHT TO EXPLANATION

Articles 13 and 14 of GDPR state that a data subject has the right to "meaningful information about the logic involved". In addition, Recital 71 states more clearly that a person who has been subject to automated decision-making:

> should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision

This requirement is clear in one aspect: the person has the right to seek an explanation of a specific decision and only *ex post*. This is important as it does not require the controller of the software to reveal the complete functionality of the system. Still, GDPR does not elucidate anywhere what constitutes an "explanation" and we will attempt to do just that.

We make two, hopefully harmless, assumptions:

© 2020 Copyright held by the owner/author(s). Published in Workshop Proceedings of the EDBT/ICDT 2020 Joint Conference, March 30-April 2, 2020 on CEUR-WS.org Distribution of this paper is permitted under the terms of the Creative Commons license CC BY 4.0.

[1] The authors of [5] claim that this right is already mandated by GDPR. The authors of [23] disagree but believe that it should be there and show how to modify the language of GDPR to do so.

CEUR-WS.org/Vol-2578/PIE1.pdf

(1) We only consider decision-making tools based on classification algorithms. Classification algorithms are "trained" on data obtained from past decisions to create a model which is then used to arrive at future decisions. It is this model that requires an explanation, not the algorithm itself (in fact, different algorithms may arrive at very similar models).
(2) We assume that the output of the algorithm is a numerical value from the range 1 to n. This covers both yes/no answers ("yes" may be then represented as the first half of the numbers and "no" as the second half) as well as categorical values (each number represent one category and we no longer assume that there is any ordering between numbers).

With these two assumptions, we can now fix the setting where we expect the right to explanation to be executed. A user submits the information about her to a decision-making tool and receives the answer X (X can be a number, a No, or a category such as "high risk"). From the wording of Recital 71 (the user has the right to challenge the decision) it is clear that the right to explanation is designed for cases where the answer the user received is different from what she expected or hoped for (say, she expected Y). The most straightforward question she may ask then is: "Why X?". When the user asks "Why X?" when she expects Y as an answer, she means in fact to ask: "Why X rather than Y?". This is the type of question that calls for a *contrastive explanation* [11]. The answer we need to provide to the user must then contain not only the explanation why the information she provided about herself generated answer X, but also what in her data has to change to generate answer Y (the one she was expecting).

When people ask "Why X?", they are looking for a cause of X. Thus, if X is a negative decision to a loan application, we would need to specify what information in their application (features used as input in the model) caused X. We also need to remember that the decision-making tool that has made a decision for the user is *replacing* a human being that used to make such decisions. In fact, a person that reports a decision to the user may not even clearly state that it is a 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 an explanation provided to her uses the language of *social attribution* [11], that is, explains the behavior of the algorithm using folk psychology. This may seem to be an excessive requirement but as we will show later in the paper, an explanation that does *not* take into account human psychology and social relations can be useless.

Last but not least, we need to be able to evaluate the quality of an explanation. This is important because – as we show in Section 3 – there are usually multiple ways of explaining X (and there are always multiple ways of explaining Y). People prefer explanations that are simpler (cite few causes) and more general (they explain more events) [9].

3 STATE OF THE ART IN MODEL EXPLANATION

Three barriers to transparency of algorithms in general are usually distinguished: (1) intentional concealment whose objective is protection of the intellectual property; (2) lack of technical literacy on the part of the users; (3) intrinsic opacity which arises by the nature of machine learning methods. Right to explanation is probably void when trade secrets are at stake (German commentary to GDPR states that explicitly [23]), but we are still left with the other two barriers. In fact, these two barriers are dependent upon each other. Complexity of machine learning methods are positively correlated with the level of technical literacy required to comprehend them. We will claim that given the current level of educational attainment in general population *and* the complexity of machine learning algorithms these barriers are insurmountable.

Let us start by putting to rest two "solutions" to this problem that have been proposed in literature on the subject. Thus, [7] suggest the following to address barrier (2):

> This kind of opacity can be attenuated with stronger education programs in computational thinking and "algorithmic literacy" and by enabling independent experts to advice those affected by algorithmic decision-making

First, even if we manage to strengthen technical literacy education (which is very unlikely given how successful we have been so far in this area) we are still left with 80% of the population which has completed their education a long time ago but may still want to use their right to explanation. Second, the cost of employing independent experts would be prohibitive and is just not feasible. (It is also not clear, how exactly these experts might be helpful.) As a solution to barrier (3), [23] suggest the following to be provided to a user as an explanation:

> Evidence regarding the weighting of features, decision tree or classification structure, and general logic of the decision-making system may be sufficient.

Indeed, this type of evidence would certainly be sufficient to understand how the system arrived at a decision. But it is completely unrealistic to expect that a layperson would be able to grasp these concepts. Anyone who taught a machine learning course at a university knows that the concepts of decision tree or neural networks are hard to grasp even for computer science majors.

In the last few years there has been very intensive work on "black-box" model explanation. Some of this work [1, 2, 10, 13, 19, 22] has been designed specifically for experts. Interpretability of a model is a key ingredient 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 for its own sake: [3, 4, 6, 15, 16, 18, 20, 24, 25]. Most of these papers are still addressed at experts with the aim of providing insights into models they create or use. In fact, only in the last three papers mentioned above, explanations were tested on human subjects and even then a certain level of sophistication was expected on their part (from the ability to interpret a graph or a bar chart to completing a grad course on machine learning). Most important though, all of these works provide explanations of certain aspects of a model (for example, showing what features influence the decision of the algorithm the most). None of them even attempts to explain fully two contrasting paths in a model leading to distinct classification results (which as we argued above is required for a contrastive explanation).

4 MODEL EXPLANATION IS HARD

We believe that explaining a black-box model of a machine learning algorithm is much harder than it is usually assumed. To make our case more vivid, we will describe our recent work [17] on designing a song recommendation system for KKBOX, Asia's leading music streaming service provider.

KKBOX has provided a training data set that consists of information of listening sessions for each unique user-song pair within a specific timeframe. The features available to the algorithm include information about the users, such as id, age, gender, etc., and about songs, such as length, genre, singer, etc. The training and the test data are selected from users' listening history in a given time period and have around 7 and 2.5 million unique user-song pairs respectively. Although the training and the test sets are split based on time and are ordered chronologically, the timestamps for train and test sets are not provided. It is worth mentioning that this structure also suffers from the cold start problem: 14.5% of the users and 26.6% of the songs in the test data do not appear in the training data.

The performance of any supervised learning model relies on two principal factors: predictive features and effective learning algorithm. Very often, these features are only implicit in the training data and the algorithm is not able to extract them itself. Feature engineering is an approach that exploits the domain knowledge of an expert to extract from the data set features that should generalize 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 our case, we created (or extracted, because they were implicitly present in data) certain statistical features such as: number of sessions per user, number of songs per each session, or the length of time a user has been registered with KKBOX. We also tried to capture the changes of user behavior over time with the following approach: for each user, we looked at how the number of songs s/he listened to per session changed over time. For that, we created two linear regression models: the first model was fitted to the number of songs per user session and the second one is fitted to the number of artists per user session. Finally, the following features were extracted from the linear models: the slope of the model, the first and last predicted values, and the difference between the first and the last predicted values.

As a result, we increased by a factor of about 10, to 185, the number of features available to the algorithm. And here is the key point: some of these derived features turned out to be extremely important in determining user's taste in songs and as a result a recommendation we provide for him or her. Yet none of these features were explicitly present in the original data! The paradox is that if someone asked us to explain how the model worked we would have had to refer to features NOT present in the data.

But this is only a part of the story. We did not use a single algorithm to make a prediction. We used five different algorithms, all of them very complex (Figure 1 shows the complexity of one of these algorithms: a simplified neural net[2] structure). Thus, here is another key point: the final model was the weighted average of all five models' predictions. It was NOT a result of one, clean algorithm.

The model that was generated by these algorithms was extremely large and complex. Since we used gradient boosting decision tree algorithms, our model was a forest of such decision trees. The forest contained over 1000 trees, each with 10-20 children at each node and at least 16 nodes deep (it took almost 128GB of RAM to derive the gradient boosting decision tree model and around 28 hours on 4 Tesla T4 GPU to create the deep neural network model).

[2]We do not explain each of these steps in detail as our point is just to show the complexity of the entire prediction process and not its technical aspects.

Now, how can a user possibly grasp this model? Assume that a user wants an explanation why song X was recommended to her rather than song Y. There will be multiple trees with the X recommendation as well as Y recommendation. Which one do we choose? These multiple trees cannot be generalized as this has been already done by the algorithm (one of the most difficult aspects of algorithms based on decision trees is optimization which is generating the simplest, most general trees). Perhaps we could generalize by approximating the answer? This approach, sometimes advocated in the literature [12], can be harmful, however. Let us say, your loan application has been rejected and you get an explanation based on an approximate model. Based on this model you are told that if your debt goes down to $10,000 you will be approved. You pay off part of your debt to satisfy that condition, apply again, and are rejected again. The reason is that the correct model we have just approximated had a $9,000 not $10,000 outstanding debt condition.

Let us summarize the obstacles in explaining the model generated by our system. A user expects a simple answer to the following question: Why did you recommend song X rather than Y?

- There may be multiple trees ("reasons") why X was recommended and similarly multiple trees why Y was not recommended. Generalizing or approximating these trees to provide a more general answer is not possible either for technical or psychological reasons.
- An explanation must refer to the actual features used by the model. Yet most of these features do not appear in the original data that describes user-song interactions. Even worse, many of them have no intuitive meaning as they are machine-generated.
- If we give up on model explanation and try instead to describe algorithms that generated the model, our task is even harder due to formidable complexity of the algorithmic design (as shown in Figure 1).

One more comment is in order. Our system used decision trees to build a model. Decisions trees are directly interpretable as each path in a tree lists simple conditions that have to be satisfied to reach a specific decision. Deep neural nets with weights attached to features and their complex interactions are *not* directly interpretable.

Complexity of machine learning models is actually worse than we described above. Machine learning is heuristics-driven and nobody expects rigorous mathematical proofs of correctness of its algorithms. What often happens is that if a model generated by some algorithm does not classify correctly the test data, a designer would stack up another algorithmic layer on top of it in hope that it improves the results. Sometimes it does but at this point nobody bothers to explain why that happened. As Ali Rahimi put it in a recent keynote talk at NIPS [14]: "Machine learning has become alchemy (…) many designers of neural nets use technology they do not really understand". If people who work with these algorithms do not understand them, how can anybody else?

5 EXPLANATIONS ARE UNNECESSARY AND CAN BE HARMFUL

We do not feel competent to answer a legal question of whether people *should* have a right to explanation. We do, however, have a few observations regarding the psychological question whether people actually want that right and will use it. We came across

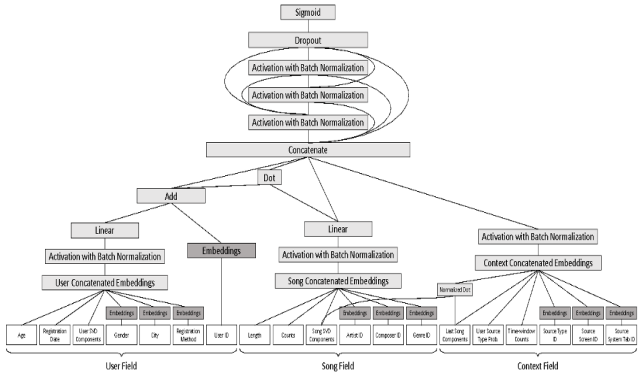

Figure 1: Structure of one of the algorithms used in the recommendation system.

three arguments routinely made to justify the requirement to explain the behavior of decision-making algorithms:

- Continuity: these decisions were previously made by humans whom we could ask for explanation. We want to keep this option.
- Gravity: these decisions (judicial sentencing, hiring, college admission) have grave consequences for our lives therefore must be justified[3].
- Trust: we trust humans more than machines therefore even if we do not always ask humans for explanations we should be able to ask machines.

We will address all three arguments with an illustrative example. Imagine you have been diagnosed with cancer and your physician suggests chemotherapy treatment. You may ask whether there are other options available to you and the doctor presents a few but still recommends chemotherapy. You may inquire further why chemotherapy is your best option to which she answers that medical studies say so. Again, you may press and ask for details of these studies but at some point (unless you are a health professional yourself) you will stop understanding her explanations. A decision - potentially a life or death decision - has been made on your behalf yet you do not insist on detailed explanation of its validity. In fact, most people will rely on the authority of the physician without asking for any explanation. One may still argue that we do not need explanations because we trust the physician (or trust her more than we trust algorithms or machines). But is it really the physician that we trust? The entire diagnostic process (MRI, X-ray, blood tests, etc.) is performed by machines, drugs used in treatment are produced by machines, and surgical procedures are performed with significant technological support. Yet we almost never ask how this technology works.

We believe that the need to get an explanation from decision-making algorithms is a simple consequence of their novelty. When we get an unexpected decision from such an algorithm, we suspect that the algorithm made a mistake and want to see the justification of the decision. In other words, we do not trust the algorithm. But we do not need an explanation to gain that trust. We believe that a much simpler and more convincing way of gaining trust is to show that the algorithm is correct and fair. Fairness and correctness can be easily verified by experts and reported back to the general population. Expert opinion is what we have used for at least 100 years in almost all of our technology, from bridge safety to GPS precision. There is no reason why it should not work here.

But there is more to our skepticism about explanations of decision-making algorithms. We believe not only that they are unnecessary, but that they can also be harmful. One unintended consequence of revealing the mechanism of an algorithm is the ability to game the system. This is unfair both to the users who did not ask for explanation (or do not have the necessary expertise to understand it) as well as the controller of the algorithm. But there is however yet another, more serious problem which has been overlooked by scholars. Algorithms are supposed to be blind to race, gender, religion etc. This blindness, however, extends to everything that is not explicitly present in the data, in particular, social context. Imagine a middle-aged woman from racial minority whose loan application has been rejected. She is told by means of an explanation that her application has been rejected primarily because of her address: she lives in socioeconomically deprived neighborhood such as Southeast LA with a documented high loan default rate among its population. But that population also happens to be mostly of the same race as the applicant (which the algorithm, of course, does not know). Needless to say, the applicant would assume that it was the race that was a hidden factor behind the negative decision. But the explanation she gets can be even more damaging. Imagine further that – as part of the contrastive explanation – the applicant is told what she should do to get an approval of the loan. To that effect, an algorithm that runs the explanation module reviews profiles in the model (these could be paths in decision trees which keep information about applicants' features such as income, type of job, age, etc.) to find the most similar ones to the applicant's. In other words, the algorithm looks for a minimal change in the applicant's profile that will give her a positive loan decision. It finds three profiles that are identical to the applicant's except for one attribute. It then suggests that the applicant should either buy a house in Beverly Hill, or increase her income by $100,000 or lower her age.

This example is not at all contrived. Every AI system is the fabled *tabula rasa*; it "knows" only as much as it has been told. A classification algorithm trained on banking data has no information about what it takes to buy a house in Beverly Hill or get a salary increase of $100,000 and it does not know that one cannot lower her age. It does not "understand" any of its own suggestions because they are generated by purely syntactic manipulation. In fact, it does not understand *anything*.

Of course, we can try to tweak the explanation module of a specific decision-making system to avoid preposterous and insulting explanations such as these. But we are rather pessimistic about the extent to which this can be done. It seems that we would need to introduce a tremendous amount of background knowledge about human behavior and social relations. This knowledge would have to be then properly organized so that the relevant part of it is easily available for a case at hand. This has been tried in the context of knowledge-base systems in the 1980s, unfortunately, without much success.

[3]Gravity of a decision does not automatically give us a right to an explanation. In most legal systems, jury verdicts are neither explained nor justified.

6 CONCLUSIONS

Black-box algorithms make decisions that affect our lives. We do not trust them because we do not know what is happening in the black-box. We want explanations. Yet, as we argued in this paper, for technical reasons, explanations at a human level are very hard to get. More than that, they can be useless or even harmful. We suggest instead, that the algorithms be analyzed from outside, by looking at their performance. Performance can be evaluated by two fundamental criteria: correctness and fairness. Machine learning community has developed reliable tests to measure algorithm correctness and we are making good progress in developing methods to test their fairness [21]. If the conclusions of this paper are correct then we need to convince policy makers (such as GDPR authors) that performance evaluation is all they can get and that it is also sufficient.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[2] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Muller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.

[3] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 598–617.

[4] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.

[5] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[6] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[7] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[8] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[9] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

[10] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 623–631.

[11] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[12] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[13] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15.

[14] Ali Rahimi. 2017. *NIPS 2017 Test-of-Time Award presentation.* Retrieved Jan 11, 2019 from https://www.youtube.com/watch?v=ORHFOnaEzPc

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. [n.d.]. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 1527–1535.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[17] Nima Shahbazi, Chahhou Mohammed, and Jarek Gryz. 2018. Truncated SVD-based Feature Engineering for Music Recommendation. In *WSDM Cup 2018 Workshop, Los Angeles*.

[18] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713* (2016).

[19] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[20] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. 1–6.

[21] Suresh Venkatasubramanian. 2019. Algorithmic Fairness: Measures, Methods and Representations. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 481–481.

[22] Marina M-C Vidovic, Nico Görnitz, Klaus-Robert Müller, Gunnar Rätsch, and Marius Kloft. 2015. Opening the black box: Revealing interpretable sequence motifs in kernel-based learning algorithms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 137–153.

[23] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 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 (2017), 76–99.

[24] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[25] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595* (2017).