UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare

Imperial College London

Paper Summary and Report

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Adam Gould

November 27, 2023

Module: 70152 - Research Tutorial - AI and Machine Learning for Healthcare

1 Paper Details

Title: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead
Author: Cynthia Rudin

Publication details: Published by Nature Machine Intelligence Intelligence, Volume 1, 2019 as a perspective article.

2 Paper Summary

2.1 Overview

In this Perspective article, Cynthia Rudin argues that using black box models and explaining their functionality after the fact is insufficient for high-stakes decision-making and instead advocates in favour of the use of interpretable machine learning despite their drawbacks [1]. Rudin's argument is built on the premise that post hoc explanations are inherently flawed, lacking detail and can complicate decisions. Despite this, black box models are more pervasive due to a wider belief that these models are inherently more accurate and better at distinguishing data patterns than interpretable models. The article challenges this belief and highlights examples of how to overcome issues with interptable models that prevent their wider adoption. The article covers a range of topics such as domain-specific definitions of 'interpretability', computationally hard problems and how comporations are incentivesed to utilise black boxes. Furthermore, potential regulation is considered that could improve the safety of systems used in high-stakes decisions. This section presents each argument Rudin makes as a summary of the ideas and examples used.

Accuracy vs Interpretability

Rudin claims that it is a pervasive 'myth' that there is always a trade-off between how accurate a model is and how interpretable a model is. To illustrate this, Rudin examines figure 1 adapted from an explainable AI report by the American defence and research organisation DARPA [2]. The graph shows a clear trend, as the effectiveness of explanations increases, the performance of the model decreases. However, no experiments were conducted to generate this graph. Quantifying the effectiveness of explanations is domain-specific and will change depending on the types of explanations provided and the intended audience. Additionally, this assumes a static data set, which is not the case in reality where data refinement techniques can lead to performance improvements and this can be more easily facilitated by an interpretable model. Furthermore, choosing which algorithms to use for such an experiment is arbitrary and could lead to a very different-looking graph.

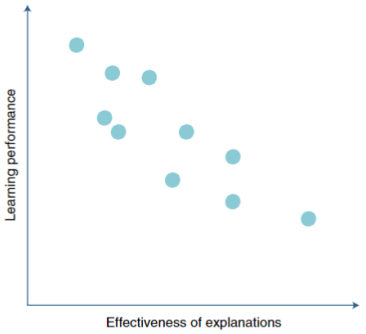

Figure 1: Trade-off representation adapted from a DARPA report [1, 2]

Rudin makes a strong counter-claim that no trade-off exists at all when using structured data with meaningful features. An example of work that Rudin conducted on predicting New York City electrical grid failures is highlighted. With this data, black box models offered very little in terms of improved performance compared to being able to refine the data set, which is better supported by transparent models.

A consequence of this belief that black box models are required for accurate performance is that academics are taught these techniques and develop tools that do not support interpretable models. One can see how this leads to further performance gains with black box models as a result of their popularity, thus perpetuating the belief of this trade-off further as shown in figure 2.

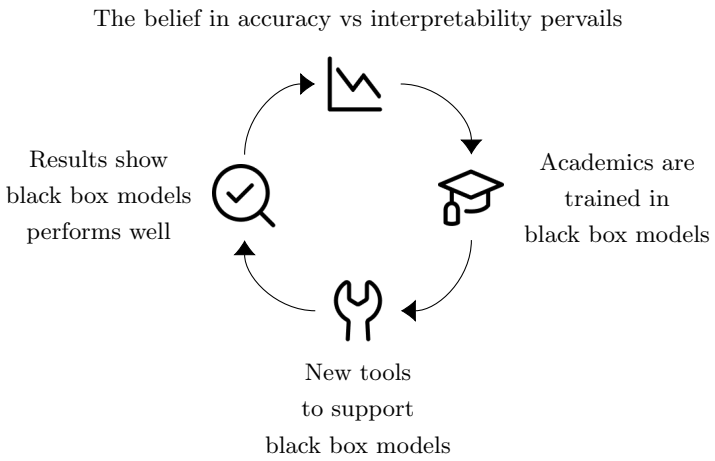

Figure 2: Accuracy-Interpretability Belief Cycle

Unfaithful Explanations

By Rudin's logic, post hoc explanations are not 100% faithful to the original model because if they were, the explanation model would be capable of making predictions indistinguishable from the black box model. The black box model would therefore be redundant as we have an interpretable model capable of doing the job. As a result, post hoc explanations must be approximations or 'summary statistics' of what the model is doing. Error can exist in this approximation and any possible errors in the explanation undermine trust in all explanations and in the original model.

Furthermore, explanations might rely on different features compared to the underlying model. The explanation model is therefore not necessarily explaining what the black box model is doing but rather trying to search for trends of the model and the data. Explanations like this are therefore ill-conceived for high-stakes decisions. The COMPAS (Correctional Offender Management Profiling for Alternative Sanction) model is used in the US criminal justice system to predict the risk of re-offenses. This is a black box model due to being proprietary. Models in this domain can utilise features that correlate with an offender's race but may not rely on race itself. As a result, ProPublica accused COMPAS of racial

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

Figure 3: Saliency Maps [1]

bias with no evidence to suggest that it uses race as a feature [3]. Rudin argues that the terminology of 'explanation' should not apply - ProPublica for instance did not create an 'explanation' of COMPAS but instead identified a 'trend' with the predictions and the data.

Explanations lacking detail

Showcasing the example of saliency maps, Rudin identifies how some explanations lack the detail necessary for understanding what a model is doing. Saliency maps highlight which regions of an image are most important when used with a CNN. However, they do not explain what the CNN is doing with this region of an image. This is emphasised by the fact that similar saliency maps can be generated for both correct and incorrect classifications, as shown in figure 3. Yet often, only explanations for correct classifications are shown which can create a false sense of trust in these explanations.

Errors and troubleshooting

Errors can occur with prediction models. With black box models, it becomes far more difficult to troubleshoot as we cannot interpret what the model is doing to lead to these errors. Circumstances in which the input data has errors lead to further confusion. Explanation models attempt to explain what the black box model is doing and utilise the input data, so often any errors will permeate through the explanations. As a result, there are now two models to troubleshoot with no clear path to do so.

Corporate incentives

Rudin goes on to argue that the reason that interpretable models do not have widespread adoption like black box models is that corporations are incentivised to keep their methods hidden. If a model is made public, or a competitor can figure out how a model operates, that will harm the profits that can be made by the model creators. Keeping a model proprietary also allows companies to make claims that are difficult to verify such as by making their model needlessly complex to appear more competitive.

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18–20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21–23 and 2–3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

Table 2 | Comparison of COMPAS and CORELS models

COMPAS	CORELS
Black box; 130+ factors; might include socio-economic info; expensive (software licence); within software used in US justice system	Full model is in Table 1; only age, priors, gender (optional); no other information; free, transparent

Figure 4: COMPAS vs CORELS [1]

The CORELS (Certifiably Optimal Rule Lists) machine learning algorithm can generate if-else rules based on data patterns. Applied to the domain of predicting criminal risk, 3 if-else rules relying solely on age, gender and criminal history can be generated that have similar accuracy to the COMPAS algorithm which relies on more than 130 features [4]. COMPAS is a proprietary algorithm, incentivised to be kept hidden from the public in order to charge its software license.

Other examples such as BreezoMeter, an ML tool that incorrectly predicted the safety of air pollution near California wildfires [5] and an example CNN used with X-ray images that relied on an area of the image tagged with the word 'portable' [6] are given. Both of these black box approaches highlight how issues with the models can go unnoticed and be dangerous to individuals. Transparent models would allow these issues to be detected and rectified and make it easier for those using the models to make decisions. However, companies that create these models do not directly suffer the consequences of incorrect predictions.

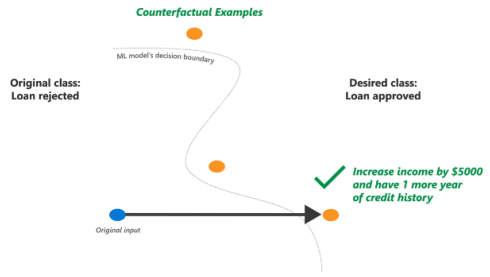

Figure 5: Counterfactual explanation [7]

Rudin considers the argument that should transparent models be used, they could be gamed. She counters this by arguing that transparency would allow those using the models to know what they need to do to change their outcome, which in many contexts such as loan applications or product ratings, could be incredibly useful. Counterfactual explanations can be used to alert people to what they need to do to change their outcome with black box models but Rudin claims that these are often insufficient because they compute the 'minimal' change necessary for a change in outcome without considering what 'minimal' means to the person it affects or the incentives of the company providing such an explanation. For example, consider an individual whose loan application has been rejected and they are told they need to have 1 more year of credit history and increase their income by $5000 as in Figure 5. This change might be infeasible for the individual and an easier change might be for them to wait 2 years. However, this may not be computed as the counterfactual nor would the company be incentivised to give this explanation if it increases their exposure.

Black Box models can uniquely uncover 'hidden patterns'

Rudin claims that interpretable models have the potential to reveal the same 'hidden patterns' that black box models can utilise but that there is greater difficulty in doing so. As a result, this leads to a belief that this ability is unique to black box models.

Significant effort to construct

Rudin recognises that there are challenges with the construction of interpretable models due to the specific tasks and contexts these models are to be deployed and the algorithmic challenges that need to be overcome.

Rudin identifies that creating optimal interpretable models such as logical models or scoring systems is a computationally hard problem. For such optimal logical models for instance, we do not just want to minimise the number of misclassifications but also the size of the model - there is no point using a model with 100 if-else rules when 3 will suffice. However, creating such a model requires solving the following optimisation problem [1]:

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[\text{training observation i is misclassified by} f]} + \lambda \times \text{size}(f) \right)$$

where n is the number of training observations, $\mathcal{F}$ is a family of logical models and λ is the classification error allowed if we were to reduce the model size by 1 unit. This is a computationally hard problem. Despite this, CORELS is able to overcome these challenges by using techniques that minimise the search space and traverse the search space faster. This example highlights how despite the existence of computational challenges, in certain contexts, with the work put in, they can be overcome.

Because interpretability cannot be easily quantified and the usefulness of an explanation changes depending on the task and the audience, creating interpretable models can be more challenging than creating black box models. Despite this, Rudin showcases their work on creating interpretable computer vision models that reason in a similar way to humans do: by identifying sections of an image that look like reference or 'prototype' images, i.e. 'this looks like that' [8]. An example of this can be seen in figure 6. The model utilises common CNNs architectures with an added prototype layer, creating a PropNet, that learns during training similarities between sections of images. During test time, the CNN identifies which section of the input image looks like learned prototypes and then uses the prototypes to classify the image.

Governance

GDPR covers a 'right to an explanation' [9] for individuals when automated processing is used. Rudin states that no guidance as to the quality of the explanation is provided so post hoc explanations, despite the flaws Rudin previously identified, could be sufficient under GDPR. To overcome this Rudin proposes two possible mandates to either enforce the creation of interpretable models by companies

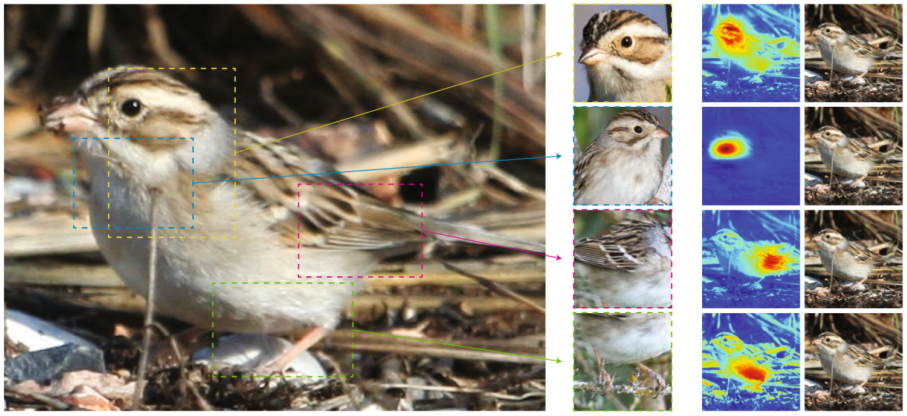

Figure 6: This looks like that model [1, 8]

when an equivalently performing interpretable model exists or failing that, enforce greater transparency in the reporting of the performance of black box models and equivalent interpretable models such that individuals can make more informed decisions about which model they should use.

Existance of interpretable models in many domains

The last argument that Rudin makes is about the existence of interpretable models in many domains. The idea behind this argument is based on the concept of Rashomon sets - sets of models with similar levels of performance for a given task. Rudin argues that for many tasks, one can expect the Rashomon set to be reasonably large, given that for instance we could train multiple neural networks with varying initial parameters on the same training data and have multiple different models each with similar performance. If a Rashomon set is sufficiently large enough, then surely at least one of the models in the set will be an interpretable model.

3 Discussion and Critique

This section discusses each point raised, critiquing both the presentation and content of each argument.

Accuracy vs Interpretability

Whilst it is accurate to claim that there is a belief that black box models are generally more accurate than interpretable ones, to claim that such a trade-off does not exist leads Rudin to fall into the same fallacy she accuses the other side of, a lack of evidence. There is no evidence to suggest that black box models always have better performance than interpretable models but, similarly, there is no evidence to say that this is not true either. The examples that Rudin uses to illustrate her point are not strong enough to counter the existence of such a trade-off, as it could be possible to find an example where, on the same data set, a black box model performs better than an interpretable one. These examples are

anecdotal and illustrate her point without supporting it. Despite this, I do not disagree with Rudin's intent, but the presentation of the argument needs to be consistent with the points made.

Unfaithful Explanations

The argument Rudin presents against post hoc explanations clearly shows that the terminology used can be misleading due to the unfaithfulness of explanations that we should not trust. A counterargument that Rudin does not explore is if there exists an acceptable level of failure in the explanations. We already accept that models will not have 100% accuracy in practice, and so we set thresholds determining when a model has acceptable performance for its use case and domain. If we are therefore willing to accept error by the classification of a model (whether it is interpretable or not), are we willing to accept a level of error in the explanation?

One argument favouring interpretable models is that if we choose to reject the explanation, we reject the prediction itself. Whereas with error in post hoc explanations of black box models, if we choose to reject the explanation, that does not directly determine if the underlying prediction should be accepted or not. So it is possible there is not an acceptable level of error in explanations and this would be an interesting concept to explore further.

Explanations lacking detail

Whilst saliency maps illustrate the point made quite well, Rudin fails to identify other explanation models that lack details, again only using anecdotal evidence. One issue with the argument put forward is that it fails to recognise that saliency maps can be used in conjunction with other methods of explaining CNNs such as occlusion tests or deconvolutional neural networks which can be used to figure out how the model behaves with areas of the image covered or which shapes are most important to name a few [10]. Combinations of explanations mean that whilst one explanation may lack detail on its own, it can be supplemented by others.

Errors and troubleshooting

The argument Rudin presents here makes logical sense, but no examples or quantifiable evidence is provided to further illustrate or support their point. Additionally, one could argue that post hoc explanations can help with troubleshooting, even if they contain errors, as for known data points we have an expectation of what the explanation should look like.

Corporate incentives

Rudin's argument about corporate incentives fails to recognise that interpretable models could be beneficial to companies. Due to the domain-specific design required for interpretable models, companies can operate as a service where they provide custom-designed interpretable models from data provided

by their clients. This works by the fact that no one-size-fits-all interpretable model has been found and even if it had been, a certain level of technical expertise would still allow companies to offer competitive services. Furthermore, offering interpretability can provide a competitive edge for models which can be a unique selling point for a company, especially given the issues Rudin has already highlighted with black box models and post hoc explanations. The argument that Rudin makes is very applicable at the current time and with the current incentives of modern companies, but an acknowledgement of possible incentives of companies in the future can show how to overcome the problems identified.

Rudin also states that 'gaming' black box models would generally be better to allow individuals to know what to do to change their outcome but there are some contexts in which transparency may be dangerous as it could allow for technology to fall into the wrong hands. This would be particularly prevalent in the context of military applications.

Additionally, whilst counterfactuals do suffer from the issues Rudin highlights, research to address these problems is in progress, looking at feasible and actionable counterfactuals [11]. Therefore, it seems amiss to discount counterfactuals altogether.

Black Box models can uniquely uncover 'hidden patterns'

Rudin's argument lacks evidence or examples to showcase interpretable models having the same ability as neural networks to identify and use hidden patterns in the data. Creating an example in the general case would be incredibly difficult, but highlighting specific instances where interpretable models have been shown to do this would emphasise her point better.

Significant effort to construct

Rudin highlights a few of the key challenges that make the construction of interpretable models incredibly difficult and showcases some clear examples that can overcome these challenges. This section of the paper offers the strongest case in favour of the use of interpretable models beyond just highlighting issues with post hoc explanations; it showcases what interpretable models offer instead.

The 'this looks like that' model is a very interesting concept showcasing how black box models such as deep learning can be leveraged to process data and identify patterns but that the reasoning system used can be easily understood and mimic human behaviour. Novel ideas such as this provide context around how challenges with the construction of interpretable models can be overcome.

Governance

Four years on from the publication of this Perspective article, society and governments are still figuring out how best to regulate the use of AI. This has become a particularly pressing matter given the prevalence of generative AI models that can now be easily accessed such as ChatGPT and Midjourney. Despite this, recent declarations and orders by governments aiming to regulate AI still fail to adequately propose concrete regulations that will reduce the risk of black box models. A recent presential Executive

Order by Joe Biden does not mention explanations, interpretability or transparency [12]. It does highlight the need for 'safe, secure and trustworthy AI'. But no definition of 'trustworthy' in the context of AI is provided - only that the National Institute of Standards and Technology will be creating standards for this. Days later at the AI Safety Summit in Bletchley Part, more than 20 nations agreed to the Bletchley Declaration. This non-binding declaration recognises the risks of AI and the need for better regulation and international cooperation but no actual regulation is put forward. Explainability is only mentioned once among a list of many concerns that need to be addressed. 'we welcome relevant international efforts to examine and address the potential impact of AI systems ... and the recognition that the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection needs to be addressed.' [13]. Rudin has highlighted a key issue and workshopped two possible solutions to mitigate AI risk, something governments are still in the process of addressing.

Existance of interpretable models in many domains

Whilst this argument makes logical sense, it relies heavily on the assumption that a sufficiently large Rashomon set does exist for a given task and even if it does, that it does contain an interpretable model. I think a better argument to be made here would be that if a domain exists in which a black box model can be applied, then an interpretable model can also be applied, given that for instance a decision tree is a universal approximator in the same way that a neural network is.

4 Summary

Overall Rudin highlights how challenges with explainable ML make it inadequate to be used for high-stakes decision-making such as in the criminal justice system or healthcare. She argues that explanations are in fact 'summary statistics' that can be unfaithful, make no sense and are difficult to troubleshoot and that as there is no difference in the performance of black box models to interpretable models, they should not be used. Instead, despite corporate incentives to make black box models and the difficulties in the construction of interpretable models, Rudin advocates for the use of interpretable models, showing that these challenges can be overcome and are worth it. Whilst Rudin's argument can at times lack sufficient evidence or valid examples to highlight the points made and does not often consider counterarguments, the perspective that Rudin presents is incredibly poignant given the risks associated with the use of AI in high stake decisions.

5 Acknowledgements

This work was supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]

References

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[2] Defense Advanced Research Projects Agency. Broad agency announcement explainable artificial intelligence (xai) darpa-baa-16-53. Technical report, DARPA, aug 10 2016. [Online; accessed 2023-11-01].

[3] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, May 2016.

[4] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(234):1–78, 2018.

[5] Michael Mcgough. How bad is sacramento's air, exactly? google results appear at odds with reality, some say. https://www.sacbee.com/news/california/fires/article216227775.html, Aug 2018.

[6] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):1–17, 11 2018.

[7] Amit Sharma. Open-source library provides explanation for machine learning through diverse counterfactuals. https://www.microsoft.com/en-us/research/blog/open-source-library-provides-explanation-for-machine-learning-through-diverse-counterfactuals/, Jan 2020.

[8] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. *This Looks like That: Deep Learning for Interpretable Image Recognition*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[9] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, September 2017.

[10] Lynn Vonder Haar, Timothy Elvira, and Omar Ochoa. An analysis of explainability methods for convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 117:105606, 2023.

[11] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, February 2020.

[12] Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/, Oct 2023.

[13] The bletchley declaration by countries attending the ai safety summit. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, Nov 2023.