Transparency, Explainability, Intellectual Property, & Copyright

Cynthia Rudin

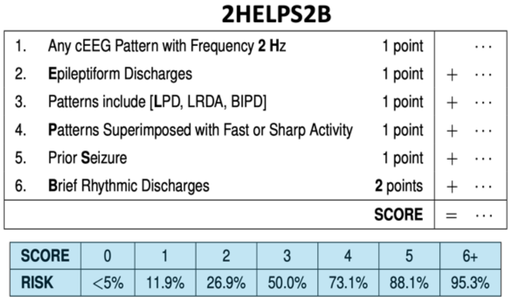
Earl D. McLean Professor of Computer Science and Engineering, Duke University

There are aspects of machine learning that are essential for regulators to understand. I will start with these and later switch to topics in transparency and copyright.

1. *(Interpretable ML For Tabular Data) While some domains benefit from ultra-complex machine learning models (e.g., computer vision, speech recognition, and language generation), many high-stakes domains (criminal justice decisions like bail and parole, many healthcare decisions) do not benefit from complex models. Very simple predictive models (small enough to fit on an index card) are just as good as deep learning for these problems.*

There are two "realms" of machine learning that behave very differently: "raw" data problems and "tabular" data problems. Raw data problems benefit from very complex models. Their prediction problems have *certain* outcomes, for instance, an image classifier should be able to determine whether an image contains a chair with more than 99% accuracy. Tabular problems are different: tabular data is what one might find on a spreadsheet, as a table of numbers. Tabular data problems predict *uncertain outcomes* such as whether someone will commit a crime after being released from prison. Tabular data problems do not benefit from complex models like deep learning. For tabular data, there are new *interpretable machine learning* algorithms that can create models that are small enough to fit on an index card yet are as accurate as deep learning.

The figure below is an example of an interpretable machine learning model, called the 2HELPS2B score. It is the only AI model used in critical care brain monitoring, and it was developed by my lab, in collaboration with neurologists. In 2HELPS2B, a critically ill patient receives points for factors that the neurologists read from measurements from the patient's brain, and the total score is translated into a predictive risk for seizure using the risk table at the bottom. 2HELPS2B was learned by an interpretable machine learning algorithm. It was trained on data from critically ill patients. It is easy for neurologists to understand, use, and troubleshoot. It is as accurate as any black box model (including deep neural networks) that anyone could construct from this patient data. The only reason why this model could be used in a high stakes decision like critical care is because humans can understand it.



2HELPS2B

1.	Any cEEG Pattern with Frequency **2 H**z	1 point	···
2.	**E**pileptiform Discharges	1 point	+ ···
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	+ ···
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	+ ···
5.	Prior **S**eizure	1 point	+ ···
6.	**B**rief Rhythmic Discharges	**2** points	+ ···
		SCORE	= ···

SCORE	0	1	2	3	4	5	6+
RISK	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

A second example comes from financial loan decisions. There was a competition in 2018 sponsored by FICO, Google, MIT, Oxford, Berkeley and others called the "Explainable Machine Learning Challenge." Entrants were given a dataset from FICO. The goal was to predict whether someone would default on a loan based on their credit history. Entrants were told to create a black box and explain it because the competition organizers did not think it was possible to create interpretable models that were accurate. Last year, new interpretable machine learning algorithms managed to produce accurate models for this dataset (as accurate as deep learning) that were also small enough to fit on an index card.

Importantly, few people understand this basic fact about machine learning problems – that interpretable models are as accurate as black box models for tabular data. Even machine learning researchers do not always understand this. This is because few AI researchers know about interpretable machine learning algorithms from within the last 20 years (though they often know about much older algorithms from the beginning of AI and the 1980's or 1990's). It is not usually taught in machine learning classes.

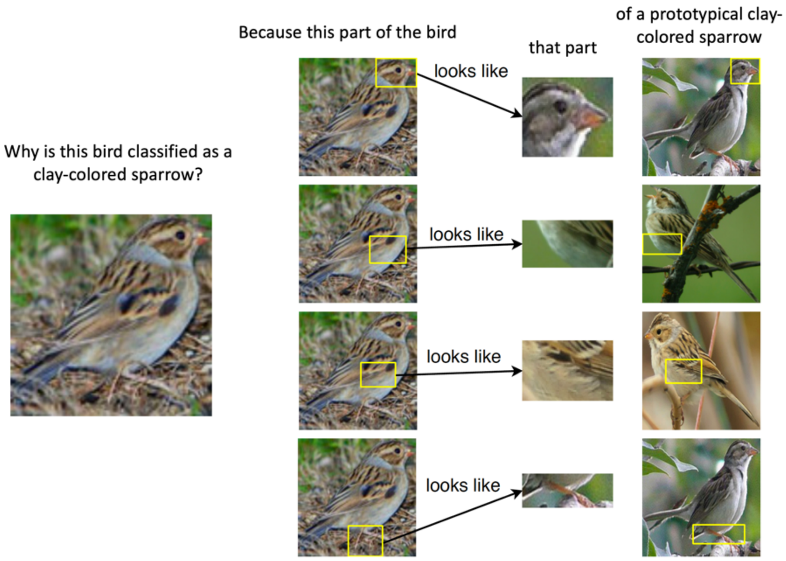
There are new mathematical proofs explaining why interpretable models are accurate in the presence of uncertain data. These proofs, and ample evidence from data from many domains, should allow regulators to aim for the following in cases with tabular data and uncertain outcomes:

NO black box models for high stakes decisions that deeply affect someone's life unless no equally accurate interpretable model can be constructed for this task.

At a minimum, this regulation should include decisions that determine freedom (bail and parole), decisions about whether someone can purchase a home (financial loan decisions) and a multitude of healthcare decisions. Ensuring that interpretable models are used in these decisions means that accountability is preserved: doctors are responsible for their own patients instead of needing to trust black box algorithms, which is important from a legal perspective. Exceptions can be made in cases where the model is 100% accurate or where humans can easily check the result, but the default should be interpretability.

2. *(Interpretable ML for Raw Data) It is often possible to create interpretable (understandable) deep learning models.*

There are some deep learning models that allow people to understand their reasoning processes. I'll include an example of how such a model works, using computer vision (specifically bird identification, but the same reasoning can work for medical imaging, face recognition, or other tasks like object recognition for self-driving cars). In the figure to the left, an interpretable deep learning algorithm called "ProtoPNet" is classifying an image of a bird (left) as a clay colored sparrow. The algorithm compares parts of the bird to parts of prototypical clay-colored sparrows and creates a numerical score for each comparison between images. The sum of scores determines the prediction from the algorithm, and in this case, the algorithm determined that the image comparisons to the clay-colored sparrows were stronger than those of other classes.


Why is this bird classified as a clay-colored sparrow?
Because this part of the bird
looks like
that part
of a prototypical clay-colored sparrow
looks like
looks like
looks like

These interpretable networks can be useful for medical imaging, self-driving cars, and facial recognition. They can allow a human to understand the reasoning process for each prediction. So far, these algorithms have gained substantial popularity, but are not nearly as popular as black box algorithms. Using them would improve the ability to troubleshoot and assign accountability in cases where the algorithm is wrong and leads to harm.

To be clear, these interpretable deep learning algorithms reveal their reasoning processes in a way humans understand. This is not the same as using a black box algorithms and trying to explain it afterwards using a "post-hoc explanation" method, discussed next.

3. *(Explanations Should Not Be Used) "Explanations" of black box model are not accurate and should not generally be trusted. It is better to use an interpretable model.*

There is a burgeoning field of machine learning called "explainable" AI (or XAI). The "explanations" from these methods are supposed to reveal what variables are being used by a black box model. However, these "explanations" often are unfaithful to the black box, particularly in medical imaging, and cannot be trusted. Tools like LIME and SHAP are not effective and untrustworthy. I have written extensively about this here:

Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead, Nature Machine Intelligence, 2019. (link to non-paywalled version of paper)

Importantly, one does not need to "explain" a black box model if one can construct an interpretable model, and as I discussed above, interpretable models are as accurate as black box models for tabular data problems, as well as computer vision problems where we can use interpretable deep neural networks.

4. *(We are owed Biometrics Transparency) Facial recognition has the potential to destroy lives and civil liberties. Voice cloning software should almost never be used. Citizens are owed the transparency over when this technology is being used on them, so their own biometrics are not used against them without permission. The government should regulate access to biometric databases and technology that uses them.*

If facial recognition is allowed to proliferate (for instance, on cell phones), it will potentially endanger anyone who enters a religious institution (mosque, synogogue), enters the witness protection program (since they could now be identified by anyone holding a cell phone), or enters a medical center (cancer center, abortion clinic). It will allow foreign governments or terrorists to track US citizens' (or anyone's) movements within the US via cell phone cameras. Voice cloning is also dangerous and software to do it should be tightly controlled and completely prohibited in the vast majority of cases.

Because of these dangers, anyone with access to a biometric dataset from which AI algorithms are built should be forced to get training and a government-issued certification. Similar to food safety, and transportation safety, AI safety is imperative. Citizens should expect transparency for when their biometrics are used for AI. We should expect it to be used to unlock our phones, by police to investigate crimes, at the border, or at large venues or sensitive facilities by licensed security teams. We should not expect it to be used for mass surveillance.

Anyone who allows even one large biometric dataset to be hacked (or released in any way) places a huge number of people in danger, and there is no possible recourse since individuals cannot change their biometrics.

5. *(Transparency in High-Stakes Decisions) If AI algorithms are used for high-stakes decisions, those algorithms should be tested carefully and their results made transparent.*

It is too easy for algorithmic designers to make mistakes or disguise poor performance if algorithms are never tested.

For instance, a federal judge ordered that the Office of the Chief Medical Examiner in New York City disclose the source code for its probabilistic genotyping software, used to analyze mixtures of DNA. As a result, a series of concerns regarding accuracy came to light, and the software was eventually discontinued. In a subsequent ruling, the judge noted that "estimates as to the likelihood of an incorrect conclusion where there actually are four or more contributors [to the DNA sample] run to over 50%." (see Brandon Garrett and Cynthia Rudin. Interpretable algorithmic forensics. PNAS, 2023.) We cannot have high stakes predictive models that make errors more than 50% of the time!

There are many cases in which models are "recycled," i.e., designed for one purpose and used for another, where they may not be as performant (e.g., bail vs. parole vs. social services). Models can easily be race- or gender-biased without detection if they are not tested.

The US government is already testing machine learning algorithms. NIST tests facial recognition algorithms, submitted by any entity (even, for instance, the Chinese government). Expanding this program to test other types of algorithms would be helpful for transparency in high-stakes decision-making.

6. *(Transparency in High-Utilization Algorithms) Recommender systems (that control social media recommendations) and other algorithms that are used by many people should be tested for 1) implications to human health and well-being and 2) disinformation, and the results should be made public.*

President Biden's executive order and Blueprint for the AI Bill of Rights makes it clear that Big Tech should no longer be allowed to trample on the public with opaque algorithms. Currently, very few academics are permitted to study these platforms. Academics who have tried to study them without permission have been punished in various ways by these companies. Since it is already clear that disinformation starts wars and causes depression in teens, I will switch topics.

7. *(Transparency in Human Health Algorithms) Health could be revolutionized by AI with help from the US government in providing test beds for important problems like heart heath monitoring from wearables.*

It is extremely difficult for academics and small companies to obtain large-scale health data, due to privacy concerns and "data hugging" by medical facilities. However, if sufficient data were available, AI algorithms could revolutionize medical imaging and disease detection and treatment.

Consider, for instance, heart monitoring. With a large fraction of the US population wearing smart watches, AI researchers could monitor human heart health and detect atrial fibrillation and other arrhythmias at a scale that has never been achieved previously. But currently, only companies that make smart watches have access to large-scale wearable device heart data. Researchers do not have access. (see How good are AI health technologies? We have no idea, with Zhicheng Guo, Cheng Ding and Xiao Hu, STAT, October 11, 2023). The only way to evaluate such AI models — or create them in the first place — is to have a large, diverse, medical dataset. The dataset must include enough patients of all kinds to ensure the AI model behaves well across different groups of people. It must be representative of all the situations in which the model might be used, whether it is in regional hospitals or major medical centers.

Since NIST already has a comprehensive evaluation program for face recognition, it could also handle evaluation of important medical AI problems such as detection of atrial fibrillation from smartwatches. That way, the benefits and flaws of proprietary algorithms (e.g., Apple Watch and FitBit atrial fibrillation detectors) can be made transparent and improved by a much broader set of researchers and companies.

8. *Copyrighted material should not be used to train machine learning algorithms without permission.*

For instance, individuals do not post photos to their own websites so that Clearview AI can scrape these photos off the internet and use them in their facial recognition algorithm to prevent entry to Madison Square Garden or Radio City Music Hall, which has actually happened. This incident illustrates two failures of AI regulation: the use of facial recognition to limit access (in a way not tied to security) to a semi-public place for which anyone can purchase a ticket, and the use of copyrighted material to train AI without permission.

9. *(Information Transparency) Generative AI is going to be increasingly difficult to detect and manage for two reasons. 1) Watermarking is almost impossible, particularly with AI generated text. 2) Many people do not care if the material is real. Thus, we should focus on provenance (i.e., providing the source of the information).*

While we should require for watermarks on AI-generated content whenever possible, it will be easy to remove those watermarks, and it will be easy for bad actors to generate content without watermarks. In other words, it will be extremely easy to circulate disinformation, bullying, and other harmful content on a massive scale. Currently, we have no way of knowing whether a phone call is spoofed (nor do we have any way of reporting the source of the spoofed calls), and we do not know whether information recommended to us on social media comes from a Russian troll farm (and no one is responsible for providing us with that information, and it is difficult to obtain).

Thus, we should create laws imposing that any entity that provides information to us at scale (e.g., our phone companies, social media) must also provide the source and provenance of that information.