INTERPRETABLEITY, HUMANS IN LOOPS, POLICIES AND POLITICS

FRANK RUDZICZ
Making Alien Minds

Think rationally  Act rationally
Think like a human  Act like a human

THE SAFETY OF AI

1. There is a risk that AI in the wrong hands, or in the hands of a select few, will perform tasks that may not be ‘globally optimal’.

2. A **bigger** risk is that AI in the right hands will:
   1. lazily be given goals that are too abstract,
   2. find a ‘trick’ to achieve those goals that we don’t understand, and
   3. result in unexpected, uninterpretable behaviour

*We need a means to explain model behaviour.*
What is actually meant by ‘explainable’?
• The wild, wild west is still working out its definitions...

Here, we will try to stick to:
• **explainable** **adj.** describes the model in general
• **interpretable** **adj.** describes a specific decision.

“the term ... holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi- mathematical way.”

Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

AMI NA AD ADI AND MOHammed BERRADA
Computer and Interdisciplinary Physics Laboratory, Sidi Mohamed Ben Abdellah University, Fez 30050, Morocco
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

FIGURE 3. Google trends result for comparing the use of “Explainable” and “Interpretable” according to the context.
When do we *want* ML to be explainable?
• We want to identify and remove bias to promote safety
• We want to leverage domain expertise
• We want to ensure generalizability and consistency
• We want to trust the system

When do we *need* ML to be explainable?
• Regulatory approval process (e.g., FDA)
• 'Right to explanation' (e.g., GDPR)
Thanks to Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, Vikas Kumar
Jean-Luc arrives at the ER.
The nurse takes age, health history, vital signs, and inputs these into a ML model.
Surprisingly, the model gives a $P(\text{admission} \mid \text{JeanLuc}) = 0.62$, which seems high.
Can we audit the system?
Let’s decompose interpretability into a few factors

TRANSPARENCY: SIMULTABILITY

• The entire model, or as much as possible, should be understood relatively holistically.

• Even basic decision trees can have thousands of nodes.

Each component should be decomposable into ‘explainable’ subcomponent.

- E.g., noun-pronoun ratio vs variance of MFCC 14’s δδ

### Regression Variables
- Age
- Gender
- Race
- Diabetic
- Smoker

### Target Variable
- Length of Stay

**Model A**

**Model B**

TRANSPARENCY: ALGORITHMIC

- Is the shape of the solution understandable?
  - Is convergence guaranteed?
    - Hill-climbing (MLE), margin maximizers (SVM), LR: **yes!**
    - Deep neural networks: **not usually**

Does it work on darker skin?

Trained with 129,450 clinical images
Tested against 2 certified dermatologists.


POST-HOC INTERPRETABILITY

• “For all we know, the processes by which we humans make decisions and those by which we explain them may be distinct.”

• “We caution against blindly embracing post-hoc notions of interpretability, especially when optimized to placate subjective demands. In such cases, one might - deliberately or not - optimize an algorithm to present misleading but plausible explanations.”

• Correlation does not imply causation.

CASE STUDY: PNEUMONIA RISK

- 14,199 pneumonia patients
  - ICD-9-CM principal diagnosis of pneumonia at admission
  - 10.86% died. Bagging is used to ‘avoid overfitting’.
  - A single 😞 70/30 train/test split is used...

- 46 features extracted, e.g.,
  - **Patient history**: chronic lung disease (+/-), admitted to ER (+/-), age (ℤ?)
  - **Physical exam**: heart rate (ℝ?), diastolic blood pressure (ℝ?)
  - **Lab findings**: potassium level (ℝ?), sodium level (ℝ?)
  - **X-rays**: pleural effusion, positive chest x-ray


**GENERALIZED ADDITIVE MODELS (GAMS)**

- Given a data set with $N$ instances, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, a standard GAM has the form
  \[
g(E[y]) = \beta_0 + \sum_j f_j(x_j)
\]
  where $g(.)$ is the link function, and “for each term $f_j$, $E[f_j] = 0$.”

- Logistic regression is a special form of GAM where each $f_j$ is linear.

- To improve accuracy, pairwise interactions can be added:
  \[
g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)
\]

CASE STUDY: PNEUMONIA RISK

<table>
<thead>
<tr>
<th>Model</th>
<th>Pneumonia</th>
<th>Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.8432</td>
<td>0.7523</td>
</tr>
<tr>
<td>GAM</td>
<td>0.8542</td>
<td>0.7795</td>
</tr>
<tr>
<td>GA²M</td>
<td>0.8576</td>
<td>0.7833</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.8460</td>
<td>0.7671</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>0.8493</td>
<td>0.7835</td>
</tr>
</tbody>
</table>

Table 2: AUC for different learning methods on the pneumonia and 30-day readmission tasks.

1.4% improvement


**CASE STUDY: PNEUMONIA RISK**

- Sort features by ‘importance’
  - Sec 5.3: ask someone fancy to rank them for you, or rank by “drop in AUC when the term is removed”
  - Better way (?): filter method, i.e., statistical tests of significance.

- Plot those features in terms of their ability to predict the outcome (risk score).
  - **Green bars** are ±1 standard deviation of the variation in the risk score (y-axis) measured by 100 rounds of bagging.

---

OK, good. Risk of pneumonia increases with age.

Uh oh, bad. Risk of pneumonia decreases if you have asthma??

It turns out, in the data, patients with a history of asthma who presented with pneumonia usually were admitted not only to the hospital but directly to the ICU.

Author’s solution: remove the term, or ask a human to redraw the graph. This assumes the channel effect (or bias) is even recognized in the first place.
CASE STUDY: PNEUMONIA RISK

- Sec 2.: “pairwise interactions are intelligible because they can be visualized as a heat map”
EXAMPLES AS EXPLANATIONS
So, Jean-Luc has been admitted as an inpatient.
The floor team now wants to decide whether he needs to go into the ICU.
Like the legal system in many jurisdictions, this decision may be based on precedent.
Can we use prior examples to interpret decisions? To explain the model?
EXAMPLES AS EXPLANATIONS

Based on slide from Shalmali Joshi
1. PROTOTYPES BY LOCAL EXAMPLES

Figure 3: Toy example to present intuition for LIME. The black-box model’s complex decision function \( f \) (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using \( f \), and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier \( f \), Number of samples \( N \)
Require: Instance \( x \), and its interpretable version \( x' \)
Require: Similarity kernel \( \pi_x \), Length of explanation \( K \)

\[ Z \leftarrow \{ \} \]

for \( i \in \{1, 2, 3, ..., N\} \) do

\[ z'_i \leftarrow \text{sample}\_\text{around}(x') \]

\[ Z \leftarrow Z \cup \{z'_i, f(z'_i), \pi_x(z'_i)\} \]

end for

\[ w \leftarrow \text{K-Lasso}(Z, K) \triangleright \text{with } z'_i \text{ as features, } f(z) \text{ as target} \]

return \( w \)
2. CRITICISMS FROM THE REAL DATA

3.2 Model Criticism

In addition to selecting prototype samples, MMD-critic characterizes the data points not well explained by the prototypes—which we call the model criticism. These data points are selected as the largest values of the witness function (5) i.e. where the similarity between the dataset and the prototypes deviate the most. Consider the cost function:

$$L(C) = \sum_{l \in C} \left| \frac{1}{n} \sum_{i \in [n]} k(x_i, x_l) - \frac{1}{m} \sum_{j \in S} k(x_j, x_l) \right|$$

(9)

- May be most useful for explaining bias in a model, instead of a decision (?)

2. PSEUDO-CRITICISMS BY SYNTHESIZING DATA

- Synthesize realistic data around decision boundaries.
- Do this along a manifold that describes realistic data.
- May also be most useful for explaining bias in a model (?)
Figure 3: We test whether ResNet models $f^1_\phi$ and $f^2_\phi$, both trained to detect hair color but on different data distributions are confounded with gender. Two samples for classifiers $f^1_\phi$ (first sub row) and $f^2_\phi$ (second sub row) are shown. The leftmost image is the original figure, followed by its reconstruction from the encoder $F_\psi$. Reconstructions are plotted as Algorithm 1 (with $\lambda = 0.01$) progresses toward crossing the decision boundary. The red bar indicates change in hair color label indicated at the top of each image along with the confidence of prediction. The label at the bottom indicates gender as predicted by $\hat{g}$. For both samples, classifier $f^1_\phi$, trained on biased data changes the gender ($1^{st}$ and $3^{rd}$ rows) while crossing the decision boundary whereas the other black-box does not.
LIVE, PIXEL-LEVEL ANNOTATIONS
Masks and Heatmaps

• So, while we got sidetracked using exemplars to explain the model itself, Jean-Luc was stabbed through the heart by a Nausicaan (or, more realistically, he took a turn for the worse).

• He needs an emergency surgery.

• In surgery, we want to identify aspects within the live video.

Warning: blood on next slide!
Which pixels are useful?
DECOMPOSABILITY – MOTIVATING EXAMPLE

First-order Taylor decomposition

\[ f(x) = f(\tilde{x}) + \left( \frac{\partial f}{\partial x} \right)_{x=\tilde{x}} \cdot (x - \tilde{x}) + \epsilon = 0 + \sum_p \frac{\partial f}{\partial x_p} \bigg|_{x=\tilde{x}} (x_p - \tilde{x}_p) + \epsilon, \]

Deep Taylor decomposition of ‘relevance’ at neuron \( j \)

\[ R_j = \left( \frac{\partial R_j}{\partial \{x_i\}} \right)_{\{\tilde{x}_i^{(j)}\}} \cdot (\{x_i\} - \{\tilde{x}_i^{(j)}\}) + \epsilon_j = \sum_i \frac{\partial R_j}{\partial \tilde{x}_i^{(j)}} (x_i - \tilde{x}_i^{(j)}) + \epsilon_j, \]
DECOMPOSABILITY

Fig. 7: Images of different ILSVRC classes ("frog", "shark", "cat", and "sheep") given as input to a deep network, and displayed next to the corresponding heatmaps. Heatmap scores are summed over all color channels of the image.
Let $A$ be a rule (set of predicates) acting on such an interpretable representation, such that $A(x)$ returns 1 if all its feature predicates are true for instance $x$. For example, in Figure 2a (top), $x$ = “This movie is not bad.”, $f(x) = \text{Positive}$, $A(x) = 1$ where $A = \{\text{“not”}, \text{“bad”}\}$. Let $D(\cdot | A)$ denote the conditional distribution when the rule $A$ applies (e.g. similar texts where “not” and “bad” are present, Figure 2a bottom). $A$ is an anchor if $A(x) = 1$ and $A$ is a sufficient condition for $f(x)$ with high probability — in our running example, if a sample $z$ from $D(z|A)$ is likely predicted as Positive (i.e. $f(x) = f(z)$). Formally $A$ is an anchor if,

$$\mathbb{E}_{D(z|A)}[\mathbbm{1}_{f(x)=f(z)}] \geq \tau, A(x) = 1.$$  

Figure 1: Sentiment predictions, LSTM
RELEVANCE MASKS

Table 3: Generated anchors for Tabular datasets
DISCHARGE
Jean-Luc had a successful heart surgery and wants to get back to his ship.

He wants to know:

- how likely he is to be re-admitted.
- is it part of a Romulan plot? A ploy to start a war?
- anything at all about his experience.
RE-ADMISSION RISK (CARUANA ET AL, SLIGHT RETURN)

• Train an extractive summarizer (‘generator’) and an encoder simultaneously.
Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems

Richard Tomsett¹  Dave Braines¹²  Dan Harborne²  Alun Preece²  Supriyo Chakraborty³

Abstract
Several researchers have argued that a machine learning system’s interpretability should be defined in relation to a specific agent or task: we should not ask if the system is interpretable, but to whom is it interpretable. We describe a model intended to help answer this question, by identifying different roles that agents can fulfill in relation to the machine learning system. We illustrate the use of our model in a variety of scenarios, exploring how an agent’s role influences its goals, and the implications for defining interpretability. Finally, we make suggestions for how our model could be useful to interpretability researchers, system developers, and regulatory bodies auditing machine learning systems.

Figure 1. Illustration of a machine learning ecosystem. Direction of arrow indicates direction of interaction (e.g., data-subjects do not interact with the system, but the system has their data).
ARTIFICIAL INTELLIGENCE FOR COMPUTATIONAL PATHOLOGY

Image interpretation plays a central role in the pathologic diagnosis of cancer. Since the late 19th century, the primary tool used by pathologists to make definitive cancer diagnoses is the microscope. Pathologists diagnose cancer by manually examining stained sections of cancer tissues to determine the cancer subtype. Pathologic diagnosis using conventional methods is labor-intensive with poor reproducibility and quality concerns. New approaches use fundamental AI research tools to build tools to make pathologic analysis more efficient, accurate, and predictive. In the 2016 Camelyon Grand Challenge for metastatic cancer detection, the top-performing entry in the competition was an AI-based computational system that achieved an error rate of 7.5%. A pathologist reviewing the same set of evaluation images achieved an error rate of 3.5%. Combining the predictions of the AI system with the pathologist lowered the error rate to down to 0.5%, representing an 85% reduction in error (see image). This example illustrates how fundamental research in AI can drive the development of high-performing computational systems that offer great potential for making pathologic diagnoses more efficient and more accurate.

SUMMARY

• By following Jean-Luc through a hospital, we’ve also visited the three main general approaches to XAI:
  • Explanations by influence of its input features
  • Explanations by examples (both actual and synthetic)
  • Explanations by heatmaps or masks

• How will (or must?!) XAI be used in practice?
REGULATION AND THE LAW
"Does your car have any idea why my car pulled it over?"
Humans are notoriously bad with information.

- Patients misread or miscommunicate their own symptoms.
- Nearly half of American adults have difficulty understanding and acting upon health information (IOM, 2004).
- Faulty memory; skill obsolescence; cognitive biases; cognitive/time limitations; recency biases; other human biases.
- Diagnoses correlate with advertising and media exposure.

Winters et al. (2012) showed that ~40,500 patients die in ICU, in the USA, each year due to misdiagnosis.

http://www.nap.edu/openbook.php?record_id=10883&page=1
Graber et al. (2005) studied one hundred cases of diagnostic error involving internists ...

- Cognitive factors contributed to 74% of cases.
  - Most common cause: ‘premature closure’.

Eddy (1990) showed top surgeons descriptions of surgical problems and asked: Should the patient have surgery?

- 50% said Yes, 50% said No.
- 40% gave conflicting answers upon retesting.


The standards that HealthCanada and the FDA used to assess software in diagnostic (Class I/Class II) devices don’t make sense anymore.

As soon as the AI makes an observation, its behaviour can change.
• The **Affordable Care Act** shifted from a fee-for-service towards a pay-for-performance model\(^1\):
  - Health IT is *rewarded*.

• Despite prohibitions in the Genetic Information Non-discrimination Act (2008), there is growing interest in using risk information for insurance stratification\(^2\):
  - Differential pricing has become one of the standard practices for data analytics vendors, introducing new avenues to perpetuate inequality.

• The (previous!) White House viewed AI as providing “increased medical efficacy, patient comfort, and less waste”\(^3\).

---


\(^3\) Bryan Biegel, & Kurose, J. F. (2016). *The National Artificial Intelligence Research and Development Strategic Plan.*

- Received in the Senate, read twice, and referred to the Committee on Health, Education, Labor, and Pensions.

Guidance I, “general wellness products”: Include “audio recordings, video games, software programs and other products that are commonly … available from retail establishments.”

- The FDA will not regulate such products as medical devices, as long as they meet two factors, specifically they:
  i) are intended for only general wellness; and ii) present low risk to users.
- These products’ value derives from information, rather than doing something directly to the body.
<table>
<thead>
<tr>
<th>Company</th>
<th>FDA Approval</th>
<th>Indication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>September 2018</td>
<td>Atrial fibrillation detection</td>
</tr>
<tr>
<td>Aidoc</td>
<td>August 2018</td>
<td>CT brain bleed diagnosis</td>
</tr>
<tr>
<td>iCAD</td>
<td>August 2018</td>
<td>Breast density via mammography</td>
</tr>
<tr>
<td>Zebra Medical</td>
<td>July 2018</td>
<td>Coronary calcium scoring</td>
</tr>
<tr>
<td>Bay Labs</td>
<td>June 2018</td>
<td>Echocardiogram EF determination</td>
</tr>
<tr>
<td>Neural Analytics</td>
<td>May 2018</td>
<td>Device for paramedic stroke diagnosis</td>
</tr>
<tr>
<td>IDx</td>
<td>April 2018</td>
<td>Diabetic retinopathy diagnosis</td>
</tr>
<tr>
<td>Icometrix</td>
<td>April 2018</td>
<td>MRI brain interpretation</td>
</tr>
<tr>
<td>Imagen</td>
<td>March 2018</td>
<td>X-ray wrist fracture diagnosis</td>
</tr>
<tr>
<td>Viz.ai</td>
<td>February 2018</td>
<td>CT stroke diagnosis</td>
</tr>
<tr>
<td>Arterys</td>
<td>February 2018</td>
<td>Liver and lung cancer (MRI, CT) diagnosis</td>
</tr>
<tr>
<td>MaxQ-AI</td>
<td>January 2018</td>
<td>CT brain bleed diagnosis</td>
</tr>
<tr>
<td>Alivecor</td>
<td>November 2017</td>
<td>Atrial fibrillation detection via Apple Watch</td>
</tr>
<tr>
<td>Arterys</td>
<td>January 2017</td>
<td>MRI heart interpretation</td>
</tr>
</tbody>
</table>
subject to the following special controls:

1. Clinical [testing] under anticipated conditions of use must demonstrate…:
   1. The ability to obtain an ECG of sufficient quality for display and analysis; and
   2. The performance characteristics of the detection algorithm as reported by sensitivity and either specificity or positive predictive value.

2. Software verification, validation, and hazard analysis must be performed. Documentation must include a characterization of the technical specifications of the software, including the detection algorithm and its inputs and outputs.


4. Human factors and usability testing must demonstrate the following:
   1. The user can correctly use the device based solely on reading the device labeling; and
   2. The user can correctly interpret the device output and understand when to seek medical care.

5. …

• Many apps serve to **shift** the **responsibility** for care and monitoring from healthcare professionals to patients themselves.
  • This may disadvantage patients who do not have the time, resources, or access to technology.
  • **What kinds of patients are favored in this new dynamic**, and might patients not well-equipped to manage and maintain their own data receive substandard care?
  • What new roles and responsibilities do the **developers** of such apps take on, and how do the ethical responsibilities of medical professionals get integrated into these differing contexts?

• How to combine **models** in different AIs? There’s no EDI in HIPAA for **models**.

Medical-record software companies are selling your health data

By Sheryl Spithoff, Special to the Star
Wed., Feb. 20, 2019

There’s a booming business in patient medical records and up to five million Ontarians are part of that boom, whether they know it or not.
The same month that GDPR came into effect, Canada issued new guidance for the Personal Information Protection and Electronic Documents Act (PIPEDA) … subsection 5(3) of PIPEDA states that “An organization may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances.” Given that consensus has not been widely achieved with regards to the details of surveillance of this type (e.g., what risks to personal information are necessary, given the technology, to achieve some perceived benefit to the person involved), it is not yet clear what a “reasonable person would consider appropriate.”
• When do we expect an explanation?
  • **Impact.** Does the action affect a 3rd party?
  • **Value.** Can something be done if we know the action was erroneous?
  • **Error.** Do we expect error?
    • **Unreliable inputs**
    • **Inexplicable outcomes**
    • **Distrust in system integrity**

• A few precedents are listed in US law.
  • Strict liability, divorce, discrimination
EU General Data Protection Regulation (enacted 2016), extends the automated decision-making rights in the 1995 Data Protection Directive to provide a right to an explanation, in Recital 71:

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

[S]uch processing 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.

Note: recitals are not binding (indeed, explainability was removed from the binding Article during the legislative process.

Solely?!
THE WAY FORWARD
Three study groups were formed within ISO/JTC1 SC 42 in 2018:

- **Computational approaches and characteristics** includes specialized AI systems (e.g., NLP or computer vision), their underlying computational approaches, architectures, and characteristics.

- **Trustworthiness** concerns approaches to establish trust in AI systems, e.g., through **transparency, verifiability, explainability, controllability**. Typical threats and risks, their mitigation techniques, and approaches to robustness, accuracy, privacy, and safety will also be investigated.

- **Use cases and applications** focuses on application domains for AI (e.g., social networks and embedded systems) and the different context of their use (e.g., health care, smart homes, autonomous cars).
STANDARDS FOR EVALUATING ML MODELS

- When comparing the performance of two or more models, several aspects must be carefully controlled and reported:
  - **Implementation** E.g., if an algorithm can be accelerated in such a way that can affect outcomes, then this must be made explicit.
  - **Hyper-parameter optimization** should not favor one model over another.
  - **Preprocessing** will not unjustly favour one model over another. E.g., removing outliers, incomplete data, or noise should not unfairly affect performance.
  - **Training and testing data** should be ecologically valid, statistically indistinct, or otherwise similar to data expected to be observed in deployment.
  - **Appropriate baselines** Any classifier should be compared against ≥1 representative, appropriate baseline. Trivial baselines should not be considered.
  - **Limiting channel effects** incl. characteristics of the manner in which data were recorded, in addition to the nature of the data themselves. Some strategies explicitly factor out channel effects.

- Appropriate statistical tests of significance must be undertaken, when possible.
We’ve talked about how AI can become safer, and how safe AI can be used to improve healthcare.

Going forward, we must leverage the advantages of our AI and human resources to save lives.

Thanks!