



XAI

e**X**plainable **A**rtificial **I**ntelligence

[Abdulrahman AlQallaf](#)

2022-12-17



Introduction – **motivation**

- **Work:** Implementing AI is hard, getting it to be used is hard.
- **Personal:**
 - Feel drawn to XAI.
 - Never liked the blind reliance on AI.



Introduction – **agenda**

1. Expose you to why it is important to have XAI.
2. General overview of the techniques.
3. A specific python package: SHAP (SHapley Additive exPlanations).



Part 1 – why is it important?

Three Real-Life Examples of AI Bias

1. Racism embedded in US healthcare

In October 2019, researchers found that an algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care heavily favored white patients over black patients. While race itself wasn't a variable used in this algorithm, another variable highly correlated to race was, which was healthcare cost history. The rationale was that cost summarizes how many healthcare needs a particular person has. For various reasons, black patients incurred lower healthcare costs than white patients with the same conditions on average.

Thankfully, researchers worked with Optum to reduce the level of bias by 80%. But had they not been interrogated in the first place, AI bias would have continued to discriminate severely.

2. COMPAS

Arguably the most notable example of AI bias is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm used in US court systems to predict the likelihood that a defendant would become a recidivist.

Due to the data that was used, the model that was chosen, and the process of creating the algorithm overall, the model predicted twice as many false positives for recidivism for black offenders (45%) than white offenders (23%).

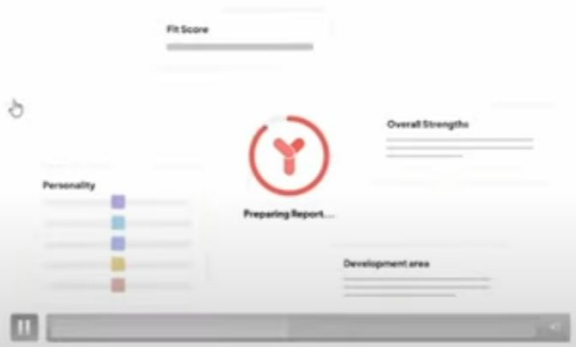
3. Amazon's hiring algorithm

Amazon's one of the largest tech giants in the world. And so, it's no surprise that they're heavy users of machine learning and artificial intelligence. In 2015, Amazon realized that their algorithm used for hiring employees was found to be biased against women. The reason for that was because the algorithm was based on the number of resumes submitted over the past ten years, and since most of the applicants were men, it was trained to favor men over women.

Part 1 – why is it important? – video interview case

What are automated video interviews?

**Just A 90
Second
Zoom Call
Generates
Up To An
18-Page
Talent
Report**



UNIVERSITY OF CAMBRIDGE
Judge Business School

From <https://www.yobs.tech.com/> archived at <https://archive.org/>

Part 1 – why is it important? – video interview case

Personality Predictions from Automated Video Interviews: Explainable or Unexplainable Models?

Dr David Stillwell
Professor of Computational Social Science
Cambridge University Judge Business School

www.linkedin.com/in/david-stillwell

Research in collaboration with Jingze Wang
(Cambridge and UCL), Franziska Leutner
(HireVue Inc. and Goldsmiths College) and
Suzanne Brink (HireVue Inc.)

Data collection

Example questions

- **Conscientiousness**
- “Tell me about a time you had a month or more to prepare for an important presentation, research project, or difficult task. Please describe the situation, how you prepared for it, and the outcome.”
- **Agreeableness**
- “Tell me about a time a close friend or coworker reached out for help while you were busy working on your own important task. Please describe the situation, how you responded and the outcome.”

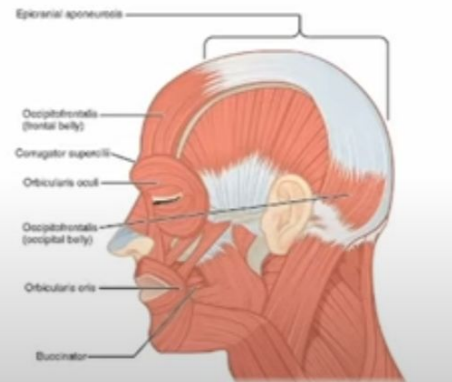
Part 1 – why is it important? – video interview case

Video features extracted

We automatically extract 1,710 features per video interview

- **Language-based**
 - Interviews automatically transcribed
 - Theory-driven word lists including LIWC, NLTK and General Inquirer
 - BERT (Bidirectional Encoder Representations from Transformer), a natural language processing pre-training algorithm.
- **Audio**
 - 50 spectral audio characteristics including rate of speech, pitch, intonation, range, gaps in speech, repetition of speech.

- **Facial expressions**
 - Labelled by Affectiva, based on the Facial Action Coding System, identifying 63 facial action units.





Part 1 – why is it important? – video interview case

Conclusions and future work

Key takeaways

- Big Five personality can be predicted from video interviews with a mean accuracy of $r=.221$
- This is comparable to humans watching video interviews who had accuracy $r=.23$ (Hickman *et al.*, 2021), and face-to-face interviews intended to assess personality $r=.28$ (Barrick *et al.*, 2000). Work colleagues also only predict each other's personality at $r=.27$ (Connelley & Ones, 2010)
- Our explainable model is less accurate at $r=.154$
- But it has the advantage of being explainable which might be key for high stakes interviews.
- The most important features are the words that the interviewee says
- Not the way that they say it, or their facial expression at the time. This is good news for practice.

Open questions

- Are the algorithms biased? And how biased are they compared to humans doing the same task?
- Are the automated personality predictions a good predictor of work behaviour and job success?
- Can prediction accuracy be increased by combining features? E.g. if the interviewee smiles while saying a certain word.

Part 1 – why is it important? – video interview case



Glass-Box



Black-Box



Part 1 – why is it important? – video interview case

“... so we did look at some of the predictions from our **explainable** Ridge **model** and we've found so it's a couple of examples so **religious language** uh which came from the German link Iowa word list um we founded that **correlated with low openness** which makes sense because there's research that finds that there's a **negative relationship between openness and religiosity** um we found that those **high in openness** use language that's more cultured intellectual **more words related to secular enlightenment** mental classes concepts ideas um uh and uh all of these kinds of language match the uh the theoretical understanding of what openness is right so high openness is meant to be **artistic interest and intellectual activities** um so it makes sense um that these kinds of language will be correlated with this type of personality.”



Part 1 – regulators push for XAI – **US**

The US government on the lookout for disability bias in employers use of AI

- Individuals with disabilities are facing unemployment rates almost twice as high as other workers who do not have disabilities... So the US government thought it was really important to focus on disability issues first.
- Using AI in:
 - Resume scanners
 - Employee monitoring software
 - **Video interviewing software** that measures a person's speech patterns or facial expressions
- Can unfairly discriminate against people with disabilities, a violation of civil rights laws

"We are sounding an alarm regarding the dangers tied to blind reliance on AI and other technologies that we are seeing increasingly used by employers," Assistant Attorney General Kristen Clarke of the department's Civil Rights Division told reporters Thursday. "The use of AI is compounding the longstanding discrimination that jobseekers with disabilities face."



Part 1 – regulators push for XAI – EU

As a case in point, the European Union's General Data Protection Regulation (**GDPR**) stipulates a right to obtain “**meaningful information about the logic involved**”—commonly interpreted as a “**right to an explanation**”¹—**for consumers affected by an automatic decision** (Parliament and Council of the European Union, 2016).²

Part 1 – regulators push for XAI – FDA

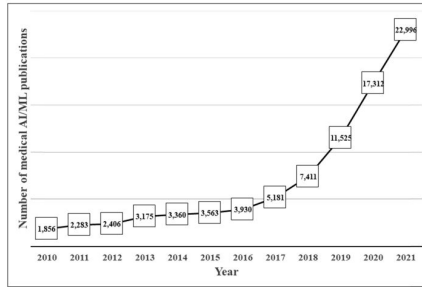


Fig. 3. The temporal growth of published medical AI/ML scientific papers.

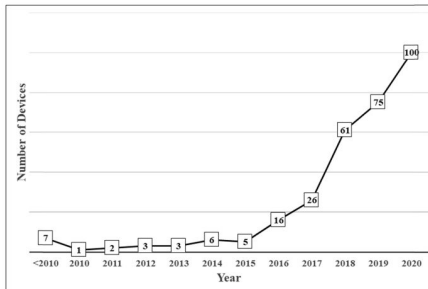


Fig. 1. The number of FDA-approved AI/ML-enabled medical devices per year.

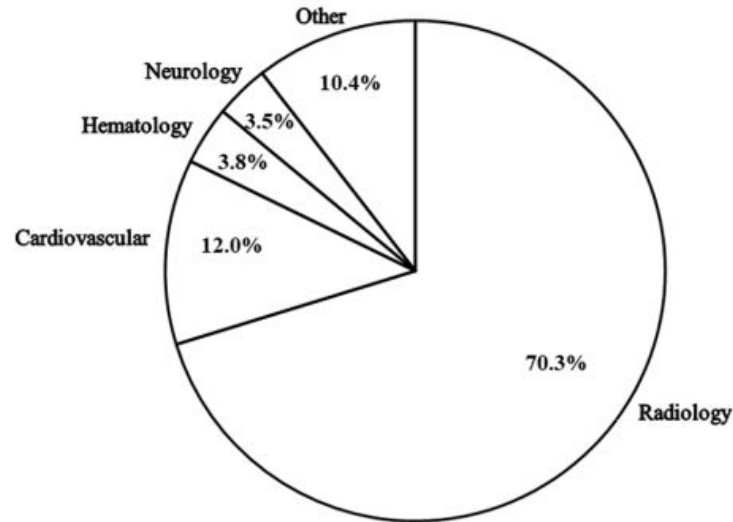


Fig. 2. Lead medical specialty review panels for the AI/ML-enabled devices.



Part 2 – XAI – a general overview



Glass-Box



Black-Box

Model **interpretability** vs. model **performance** trade-off:

- **Simple models** are interpretable (e.g., decision tree), but usually have problems in performance.
- **Complex models** usually yield better performance, but it's difficult to interpret their results (e.g., neural networks).

→ Tendency from DS to fixate on model performance.

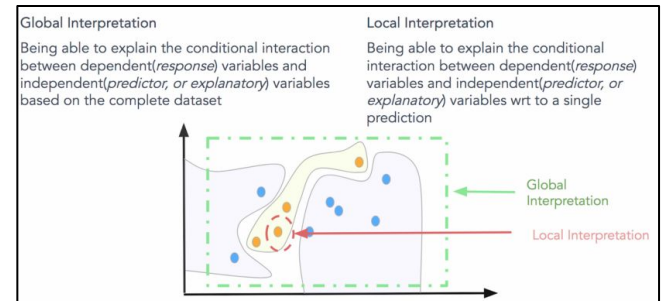
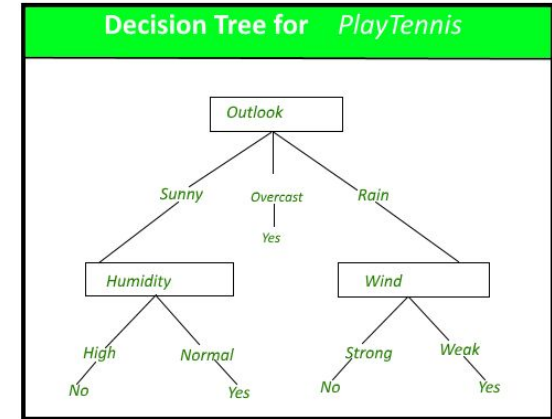


Part 2 – XAI – **desirable features**

- **Fairness:**
 - What drives model predictions?
- **Accountability:**
 - Why did the model take a certain decision?
- **Transparency:**
 - How can we trust model predictions?

Part 2 – XAI – methods categorization

- **Intrinsic vs. post hoc**
- **Model specific vs. model agnostic**
- **Local vs. global**





Part 2 – XAI – traditional method – **EDA**

→ switching to Jupyter ...

Part 2 – XAI – cool method – **Surrogate Model**



Part 2 – XAI – cool method – **Surrogate Model**

1. Choose a dataset This could be the same dataset that was used for training the black box model or a new dataset from the same distribution. You could even choose a subset of the data or a grid of points, depending on your application.
2. For the chosen dataset, get the predictions of your base black box model.
3. Choose an interpretable surrogate model (linear model, decision tree, ...).
4. Train the interpretable model on the dataset and its predictions.
5. Congratulations! You now have a surrogate model.
6. Measure how well the surrogate model replicates the prediction of the black box model.
7. Interpret / visualize the surrogate model.

Sample illustration of a surrogate tree model



Part 3 – **SHAP**

→ switching to [Github](#) ...



Questions?

References:

- <https://datatron.com/real-life-examples-of-discriminating-artificial-intelligence/>
- <https://www.youtube.com/watch?v=V-48D7giYik&list=PPSV>
- <https://interpret.ml/>
- <https://www.npr.org/2022/05/12/1098601458/artificial-intelligence-job-discrimination-disabilities>
- <https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1391>
- https://scholarlycommons.henryford.com/cgi/viewcontent.cgi?article=1359&context=radiationoncology_articles
- <https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476>
- <https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739>
- <https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608>
- <https://github.com/slundberg/shap>
- <https://shap.readthedocs.io/en/latest/>