



# EXPLAINABILITY + RESPONSIBILITY DESIGNING TRUSTWORTHY AI MODELS

Dr Stephanie Baker  
James Cook University

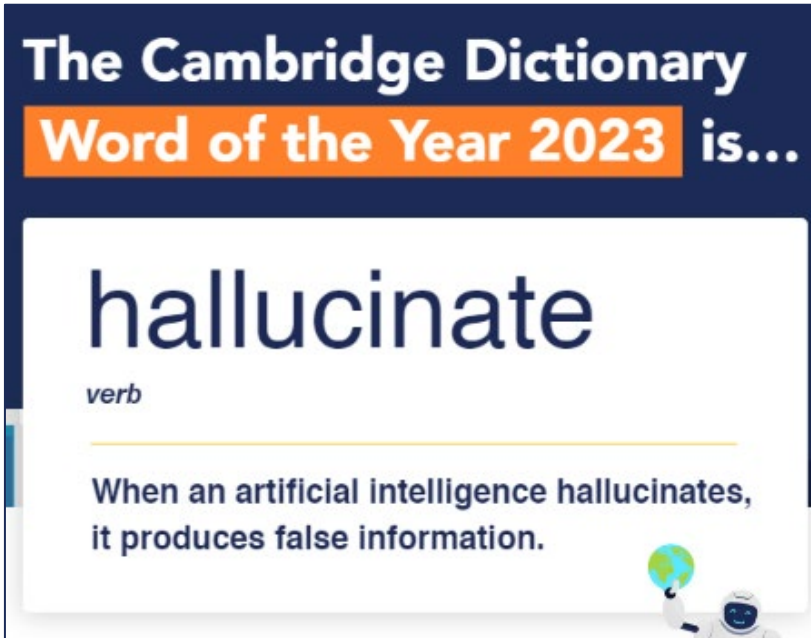
E: [stephanie.baker@jcu.edu.au](mailto:stephanie.baker@jcu.edu.au)  
P: +61 7 4232 1561

# THE COLLINS WORD OF THE YEAR 2023 IS...

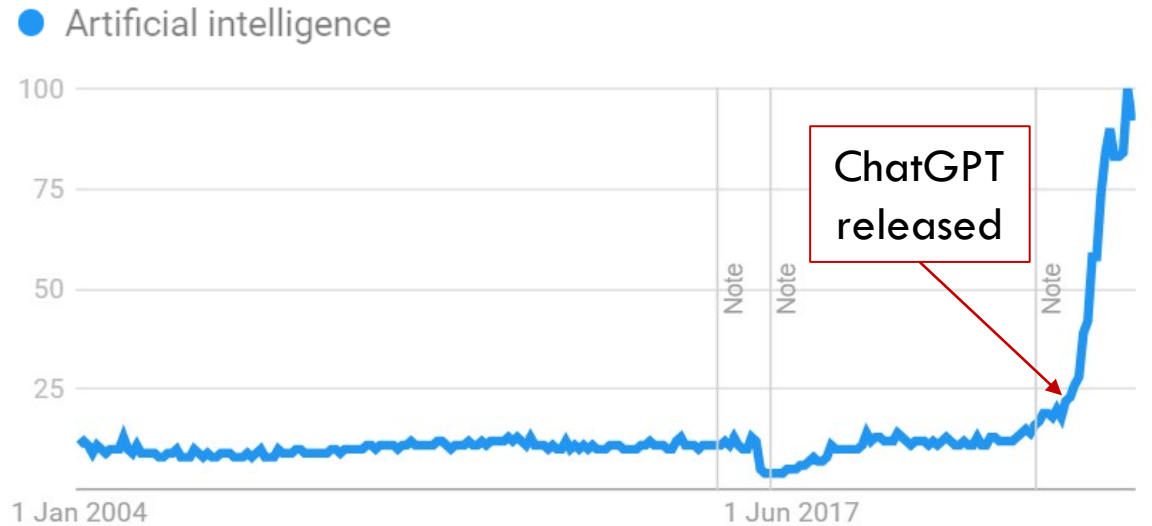
## AI

'AI', a term that describes 'the modelling of human mental functions by computer programs' has been named Collins Word of the Year 2023. Considered to be the next great technological revolution, AI has seen rapid development and has been much talked about in 2023.

**Source:** Collins Dictionary, "The Collins Word of the Year is...", <https://www.collinsdictionary.com/woty> (accessed Nov. 22, 2023).



**Source:** Cambridge Dictionary, "The Cambridge Dictionary Word of the Year 2023 is...", <https://dictionary.cambridge.org/editorial/woty> (accessed Nov. 22, 2023).



Worldwide. 01/01/2004 - 22/11/2023. Web Search.

**Source:** Google, "Google Trends," 2023. <https://trends.google.com/trends/> (accessed Nov. 22, 2023).

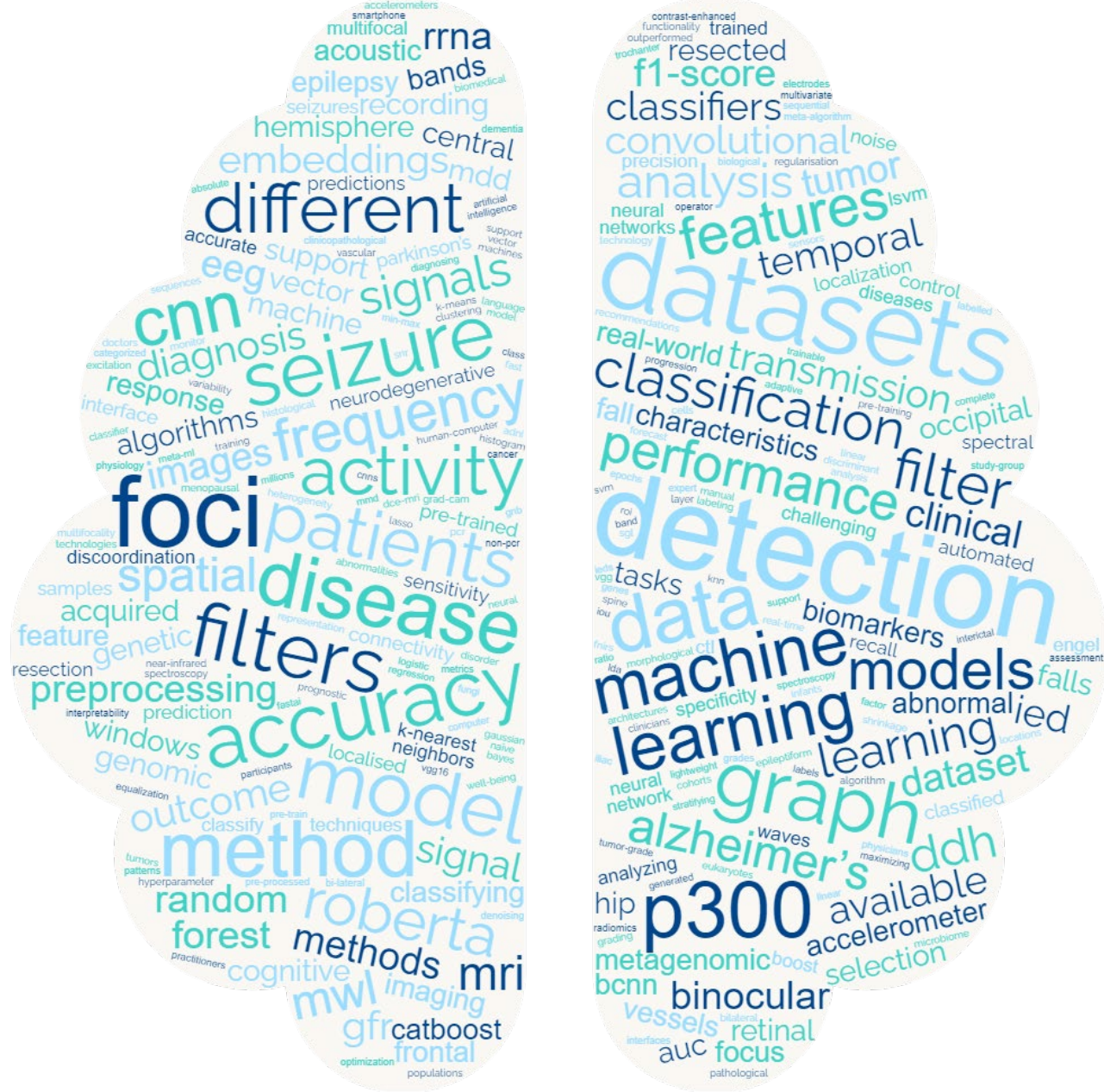
77%

of papers

89%

of posters

in IEEE SPMB 2023 include machine learning or related concepts



Word cloud generated from IEEE SPMB 2023 abstracts using <https://www.wordclouds.com/>



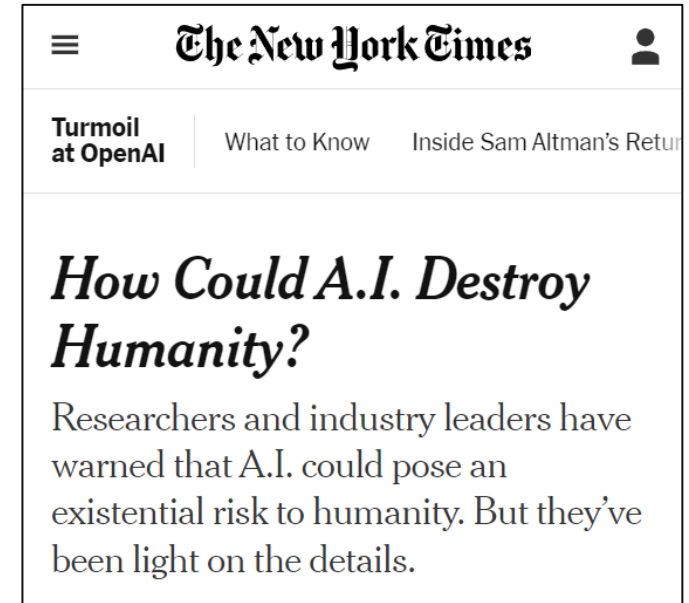
The screenshot shows the BBC News website. At the top, there is a navigation bar with the BBC logo, a search icon, and a menu icon. Below this is a red banner with the word "NEWS" in white and a "Menu" button. The main content area is titled "Tech" and features a large headline: "Artificial intelligence could lead to extinction, experts warn". Below the headline, it says "30 May" and "Comments".

C. Vallance, "Artificial intelligence could lead to extinction, experts warn," BBC, May. 30, 2023.



The screenshot shows the The Guardian website. At the top, there is a dark blue banner with the text "Support the Guardian" and "Fund independent journalism with \$5 per month". Below this is a yellow "Support us" button. The main navigation bar includes "News", "Opinion", "Sport", "Culture", and "Lifestyle". The main content area is titled "Artificial intelligence (AI)" and features a large headline: "Five ways AI might destroy the world: 'Everyone on Earth could fall over dead in the same second'". Below the headline, it says "Artificial intelligence is already advancing at a worrying pace. What if we don't slam on the brakes? Experts explain what keeps them up at night".

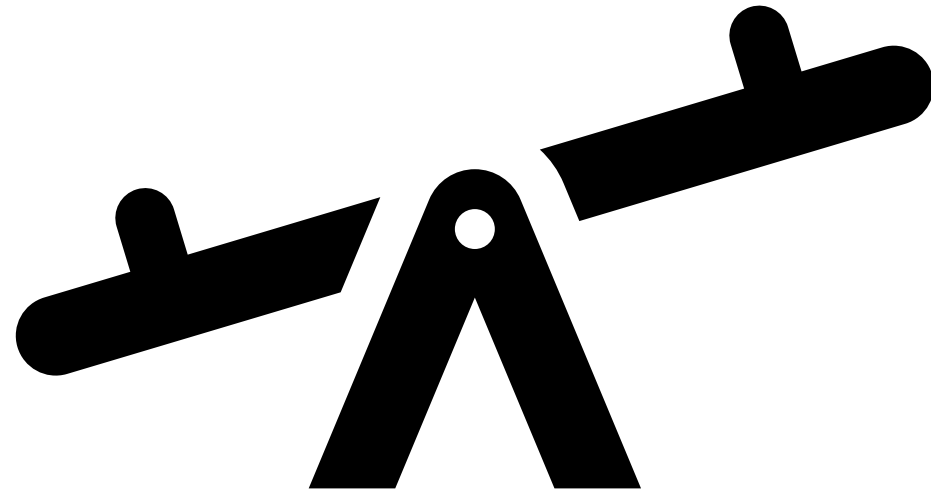
S. Rose, "Five ways AI might destroy the world: 'Everyone on Earth could fall over dead in the same second,'" *The Guardian*, Jul. 7, 2023



The screenshot shows the The New York Times website. At the top, there is a navigation bar with the text "The New York Times" and a search icon. Below this is a white banner with the text "Turmoil at OpenAI" and "What to Know Inside Sam Altman's Return". The main content area features a large headline: "How Could A.I. Destroy Humanity?". Below the headline, it says "Researchers and industry leaders have warned that A.I. could pose an existential risk to humanity. But they've been light on the details."

C. Metz, "How Could AI Destroy Humanity?," *The New York Times*, Jun. 10, 2023

# How do we balance the benefits of AI with the risks?



## **Responsible AI:**

- Application of ethics to AI development
- Development of AI models that benefit humanity

## **Explainable AI:**

- Tools to make AI models more transparent and interpretable to developers and end users

**Responsibility requires  
explainability**





# DEEP DIVE: RESPONSIBLE AI

# Why does responsible AI matter?

- If medical, surgical, or other healthcare AI systems fail, significant harm can be caused – up to and including loss of life
- AIs can learn from biased data
  - Example: AI diagnosis systems can fail to diagnose conditions in minority groups due to historical underrepresentation in medical data [1]
- AIs can share outdated, misleading, incorrect, or harmful information
  - Example: ChatGPT has been shown to give inadequate information about medical conditions [2] and misleading information about mental health disorders [3]
  - Example: AI chatbots can respond inappropriately to mental health crises [4, 5]

[1] I. Straw, “The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future,” *Artif. Intell. Med.*, vol. 110, p. 101965, 2020

[2] E. A. M. van Dis *et al.*, “ChatGPT: five priorities for research,” *Nature*, vol. 614, no. 7947, pp. 224–226, 2023

[3] Y. H. Yeo *et al.*, “Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma,” *Clin. Mol. Hepatol.*, vol. 29, no. 3, pp. 721–732, 2023

[4] L. Martinengo *et al.*, “Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis,” *J. Affect. Disord.*, vol. 319, pp. 598–607, 2022

[5] L. Laestadius *et al.*, “Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika,” *New Media Soc.*, Dec. 2022



# Why do we need responsible AI?

- We want AIs that act to benefit people, environments, and societies
- Medical AIs, like medical doctors, should strive to “do no harm”

# What is responsible AI?

- It depends who you ask! There are many differing frameworks and guidelines.
- Industry frameworks include:
  - Meta: <https://ai.meta.com/responsible-ai/>
  - Microsoft: <https://www.microsoft.com/en-us/ai/principles-and-approach>
  - Amazon: <https://aws.amazon.com/machine-learning/responsible-ai/>
  - OpenAI: <https://openai.com/safety> (and hyperlinks within this page)
  - Google: <https://ai.google/responsibility/responsible-ai-practices/>
  - Intel: <https://www.intel.com/content/www/us/en/artificial-intelligence/responsible-ai.html>
  - IBM: <https://www.ibm.com/impact/ai-ethics>
- Government agency frameworks/policies include:
  - Department of Industry, Science and Resources (Australia): <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles>
  - India AI: <https://indiaai.gov.in/responsible-ai/homepage>
  - National Institute of Standards and Technology (United States): <https://www.nist.gov/trustworthy-and-responsible-ai>

# What is responsible AI?

- All of the frameworks, policies, etc. have a few things in common
- They all **seek to provide a set of principles or guidelines focused on how responsible AI models should be developed, deployed, and used in order to meet ethical (and sometimes legal) standards.**
- Responsible AI is sometimes used interchangeably with “fair AI”, “ethical AI”, “trustworthy AI”, etc. – but the literature shows that responsible AI should be all of these things and *more*.

privacy

robustness

transparency

data protection

resilience

data transparency

data security

prevent failures

provide explanations

avoid bias

governance

auditability

protection during failure

diversity & inclusion

contestability

minimise harm

fairness

accountability

safety

privacy

robustness

transparency

data protection

resilience

data transparency

data security

prevent failures

provide explanations

avoid bias

governance

auditability

protection during failure

diversity & inclusion

contestability

minimise harm

fairness

accountability

safety

# Responsible AI principle #1: Privacy

- AIs must preserve the privacy of data used in development; AI should not reveal sensitive data through reverse engineering or malicious attacks.
- All government and academic sources agree that privacy is essential.
- Most industry bodies have put some thought towards privacy.
- Data security and data protection are often mentioned alongside privacy.

## Responsible AI principle #2: Robustness

- AI models should be resilient against failure and malicious attacks and operate as intended across all use cases.
- Almost all sources consider robustness or a related phrase such as reliability or resilience.

## Responsible AI principle #3: Transparency

- AI models should be able to describe or explain their decisions in a way that is meaningful to stakeholders, from developers to end-users.
- Strongly tied to explainability and the field of explainable AI.
- Almost all sources agree that transparency is essential.



## Responsible AI principle #4: Fairness

- AI models should be inclusive, treat people and scenarios fairly, and not discriminate.
- All sources agree that fairness is important.
- Some sources also mention diversity and/or inclusion, either as part of fairness or as an additional principle.

## Responsible AI principle #5: Accountability

- AI models should be able to justify their decisions to the satisfaction of stakeholders from non-technical users to governing bodies.
- AI models should have mechanisms for stakeholders to question, refute, or give feedback on AI decisions, with processes for timely resolution.
- Accountability is considered by all sources.
- Disagreement on who should be held accountable: the AI or the developers?
- Several sources identified that transparency supports accountability.

## Responsible AI principle #6: Safety

- AI models should not cause harm to people, environments, or societies.
- Discussed by all academic and government sources, and most industry sources.
- Several sources group safety with concepts such as robustness and reliability.

# Six principles of responsibility

1. Privacy
  - Which principles do you think are the most important?
2. Robustness
  - Are there any other principles that you think are important to responsible AI?
3. Transparency
  - Are there any principles I've grouped together that you would consider separate?
4. Fairness
  - Tell me using this link or the QR code (*to be provided during live talk*)
5. Accountability
6. Safety

Responsible  
AI



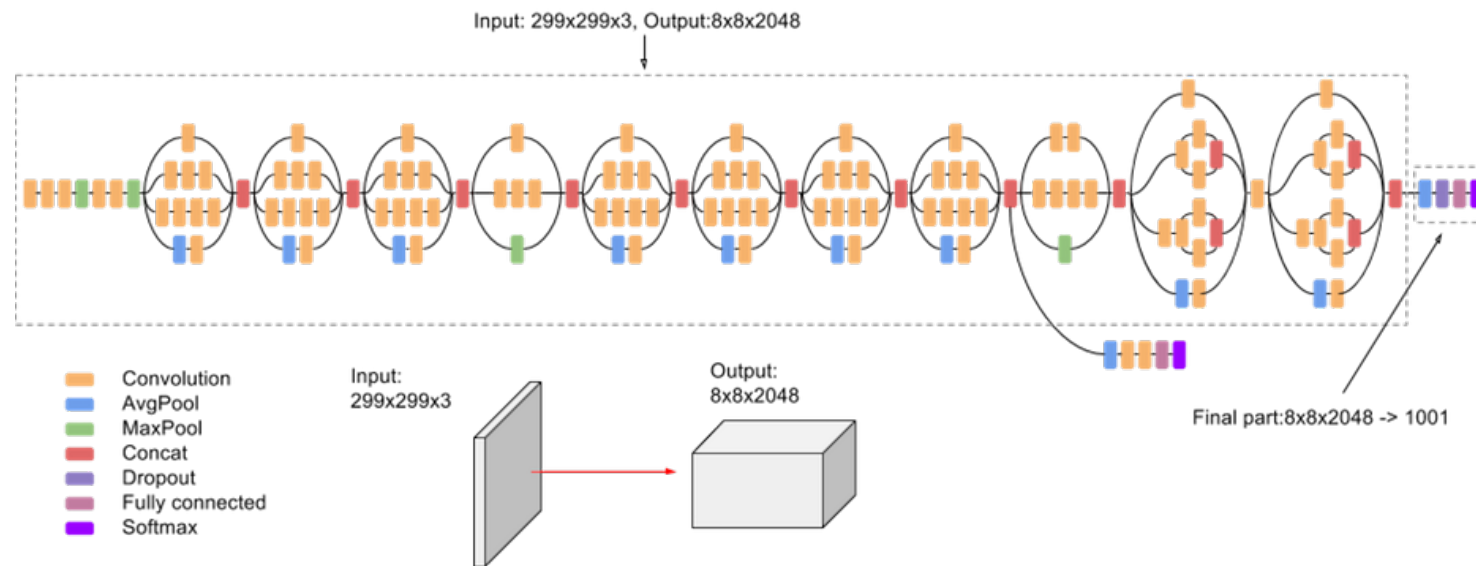
Explainable  
AI



# DEEP DIVE: EXPLAINABLE AI

# Explainable AI (XAI)

- AI models have become increasingly complex, making each decision difficult to understand



Inception v3 architecture. Source: Google Cloud, “Advanced guide to Inception v3,” 2023.  
<https://cloud.google.com/tpu/docs/inception-v3-advanced> (accessed Nov. 20, 2023).

# Explainable AI (XAI)

- AI models have become increasingly complex, making each decision difficult to understand
- XAI focuses on making models and their decisions interpretable to users, developers, and other stakeholders
- The level of explanation required will vary depending on the stakeholders and the application
- Two types of XAI: Explainable by design, and post-hoc explanations

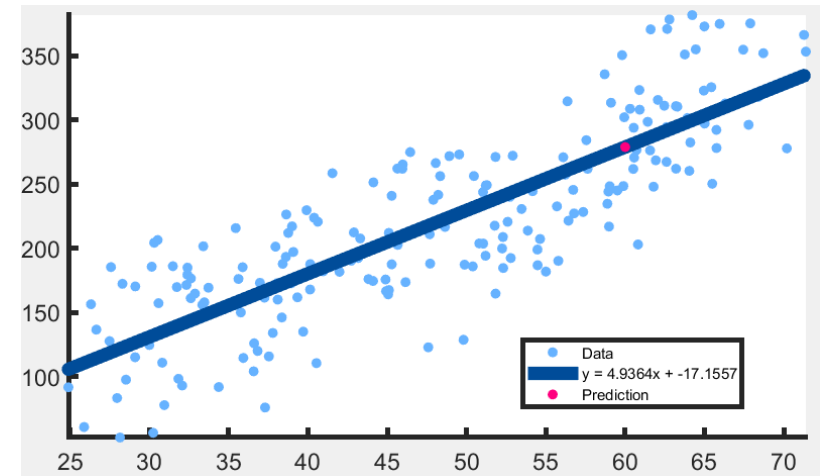


# Explainable by design

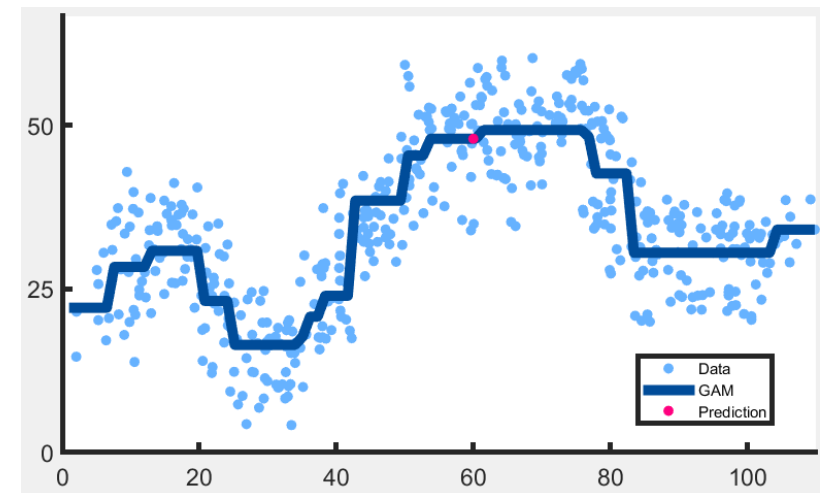
- Many early “AI” models are explainable due their simplicity; their design and decision mechanisms can easily be inspected and understood.
- Examples: linear regression, logistic regression, decision trees, clustering, k-nearest neighbour, etc.

# Explainable by design: line fitting approaches

- Line fitting approaches can be easily interpreted by inspecting the data and the line.
- Some mechanisms for generating lines are more straightforward than others.
- New predictions can be readily explained by showing the new prediction along the line.

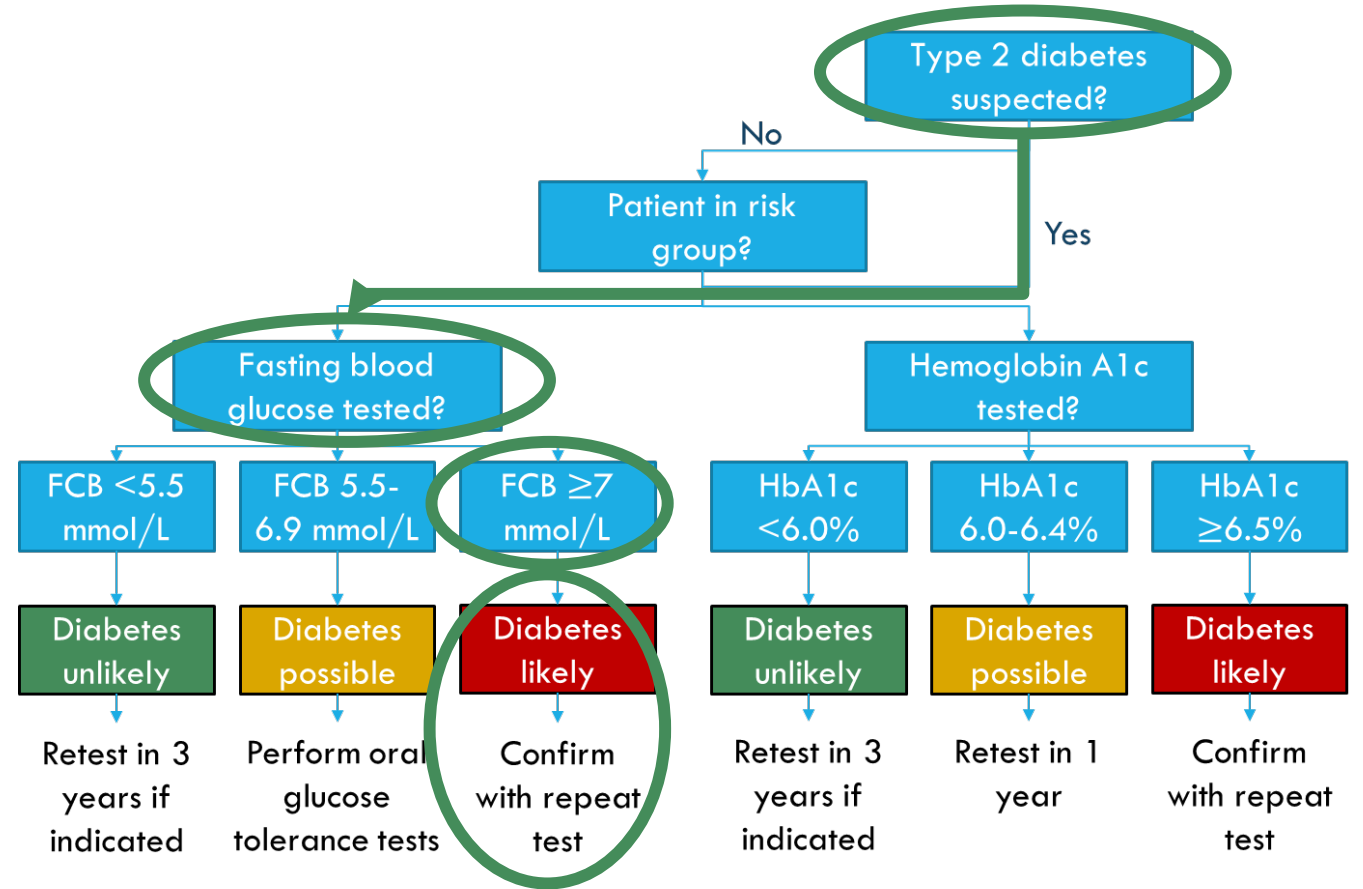


Linear regression (above) and generalized additive models (below)



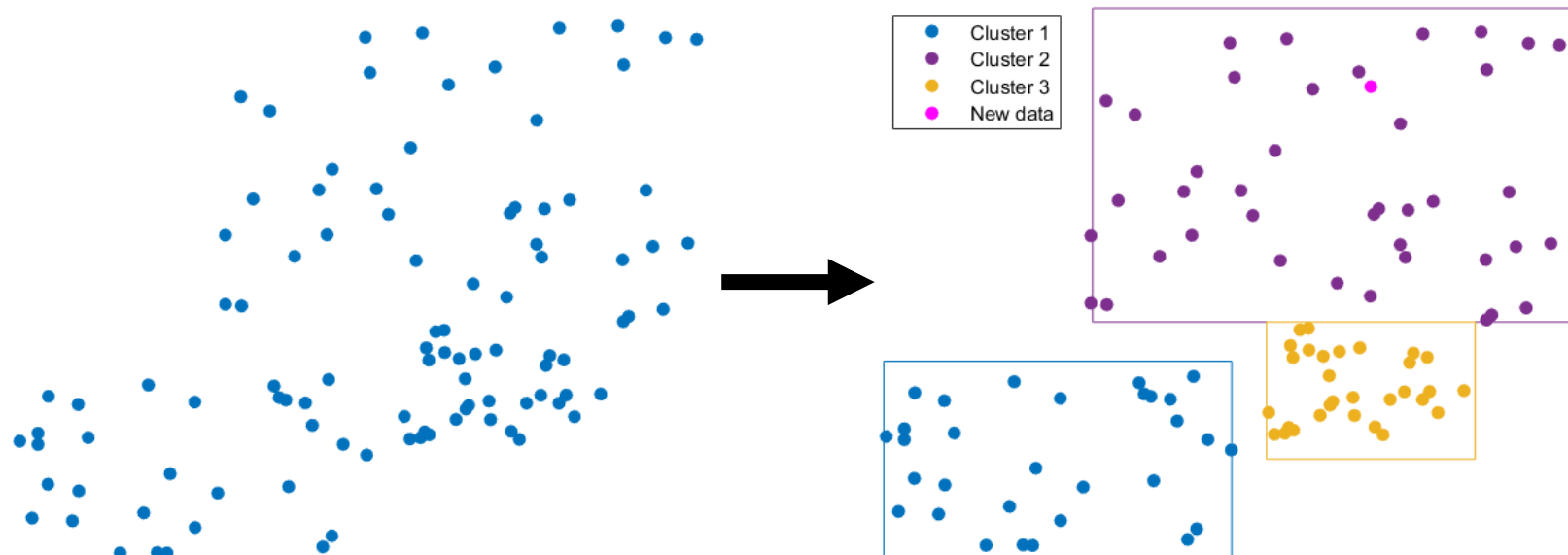
# Explainable by design: decision trees

- Decisions follow a logical path that can be viewed and readily interpreted.
- Approaches that utilise multiple decision trees are more common in the literature (e.g. random forest).
- Interpretability decreases as more trees are added.



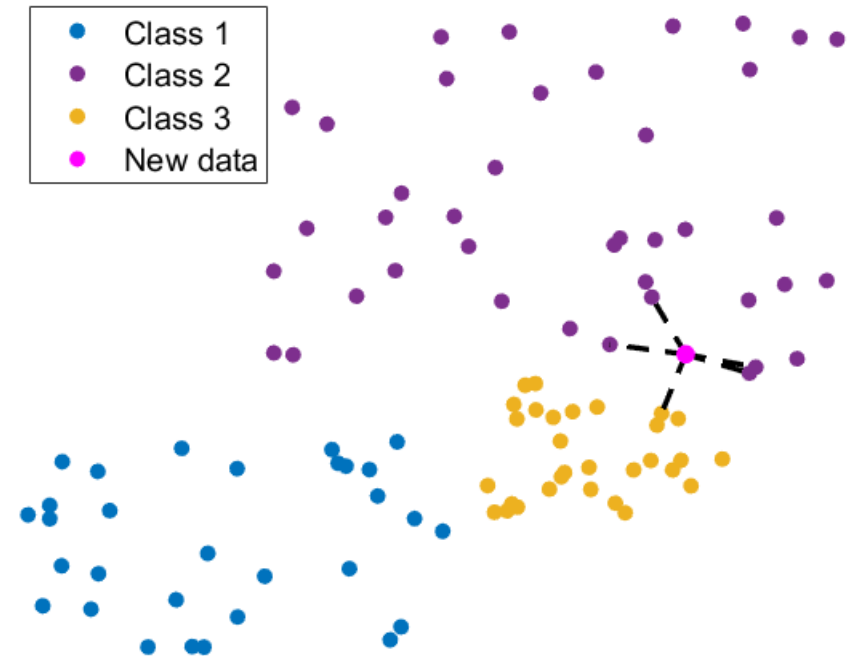
# Explainable by design: clustering approaches

- Clustering approaches can be easily visualised by inspecting data and clusters.
- Some methods of generating clusters are more interpretable than others.
- New predictions can be readily interpreted by showing the new prediction in the cluster.



# Explainable by design: k-nearest neighbours

- Prediction for a new data point is based on the value of the k-nearest neighbours
- Easy to visualise for small values of k, becomes more cluttered as k increases



# Explainable by design

- Simplicity is good for interpretability.
- But there's a reason these models aren't seen much in modern literature – their simplicity often fails to capture complex features.

# Post-hoc explainability

- Methods for explaining models and decisions after they are made.
- Useful for complex models that are difficult to visualise.
- Local post-hoc explanations seek to explain individual predictions.
- Global post-hoc explanations seek to explain the overall workings of the model.
- Quite a few different methods, we'll explore just a few.

# Post-hoc explainability: feature importance approaches (global)

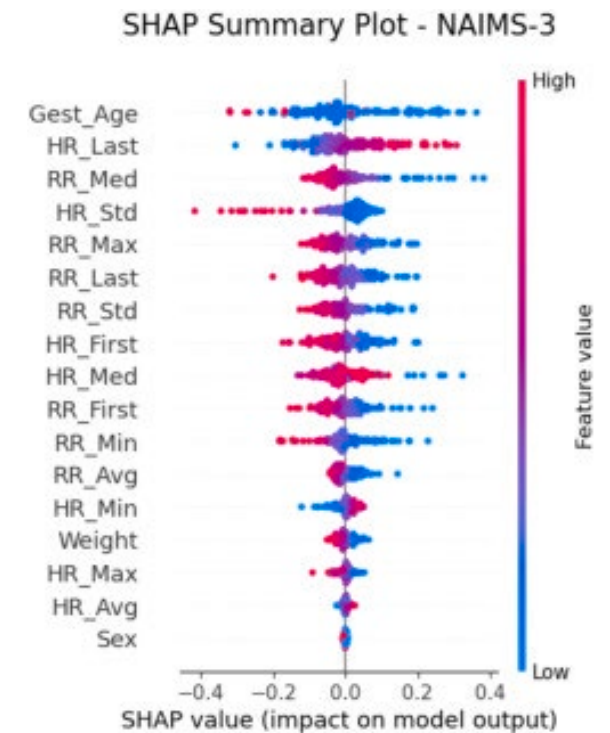
- **How can it help your research?** Biomarker discovery, improve your model by revealing issues, explain your results to users.
- Several methods have been proposed in the literature to rank the importance of features to a model.
- Methods include permutation importance, SHapley Additive exPlanations (SHAP), individual conditional expectation (ICE) plots, and more.



# Post-hoc explainability: feature importance approaches (global)

## SHAP [1]

- Game-theory approach that quantifies the impact that each feature ('player') has on the models' predictions ('games').
- SHAP values are calculated for an individual prediction; positive values push the output higher, negative values push the output lower.
- Global SHAP is calculated by aggregating results for all (or a subset of) the individual predictions.
- Can be applied to any model.



(a) NAIMS-3

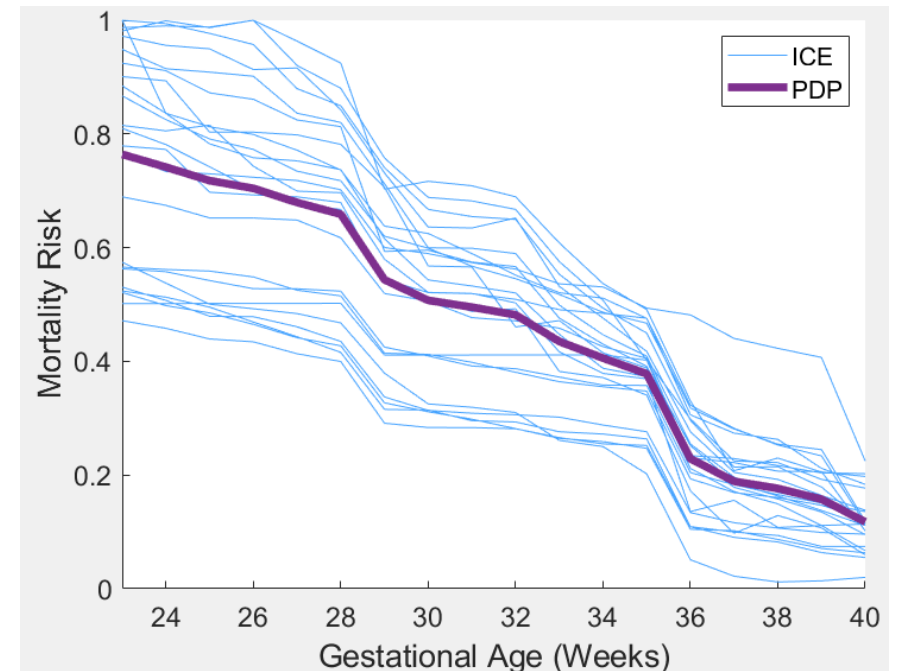
[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

Figure Source: S. Baker, et al., "Hybridized neural networks for non-invasive and continuous mortality risk assessment in neonates," *Comput. Biol. Med.*, vol. 134, p. 104521, 2021.

# Post-hoc explainability: feature importance approaches (global)

## ICE & PDP plots

- Both types of plot look to evaluate the impact of changing a single feature on the models' predictions.
- Individual conditional exception (ICE) plots show the impact of changing a feature on each individual prediction in the dataset.
- Partial dependence plots (PDP) show the overall impact of changing a feature on the model, by averaging the ICE plots.
- Can be applied to any model.



# Post-hoc explainability: feature importance approaches (global)

## Permutation importance

- After model training, all values for a particular feature are shuffled to assess the impact of a given feature on predictions.
- Features that have the most impact on predictions will have the highest permutation importances.
- Can be applied to any model.

Feature	Weight
Gestational Age	+0.541
Most recent heart rate	+0.335
...	...
Sex	+0.029

# Post-hoc explainability: feature importance approaches (local)

- **How can it help your research?** Analyse predictions that your model got wrong, explain your model to end users.
- Many out there, SHAP and Local Interpretable Model-Agnostic Explanations (LIME) are perhaps the most popular.

# Post-hoc explainability: feature importance approaches (local)

## SHAP (again)

- SHAP values are calculated for an individual prediction; positive values push the output higher, negative values push the output lower.
- Can be applied to any model.

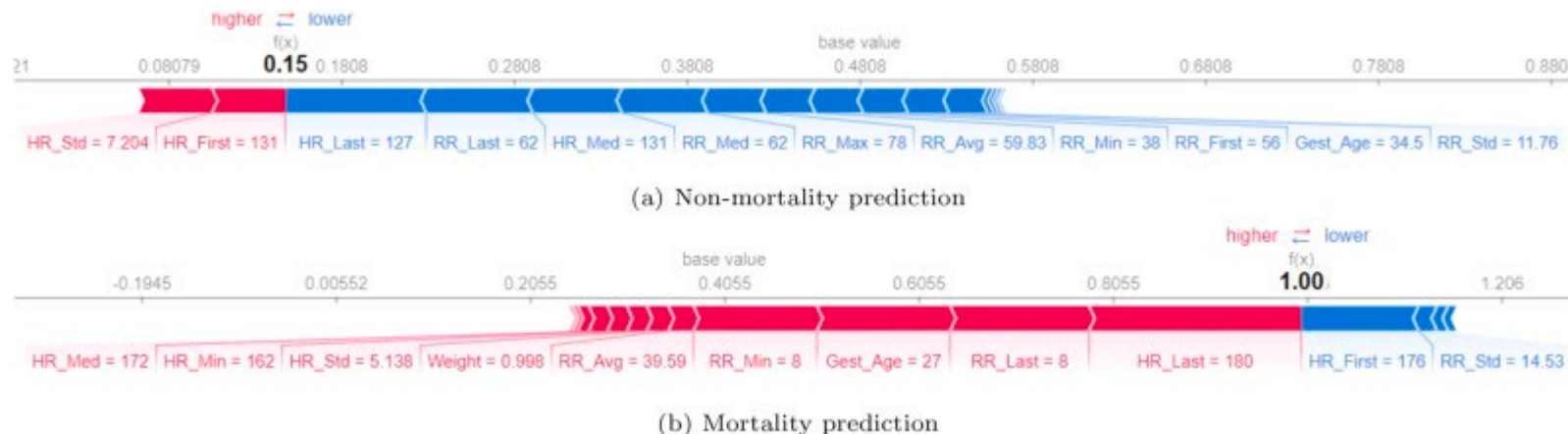


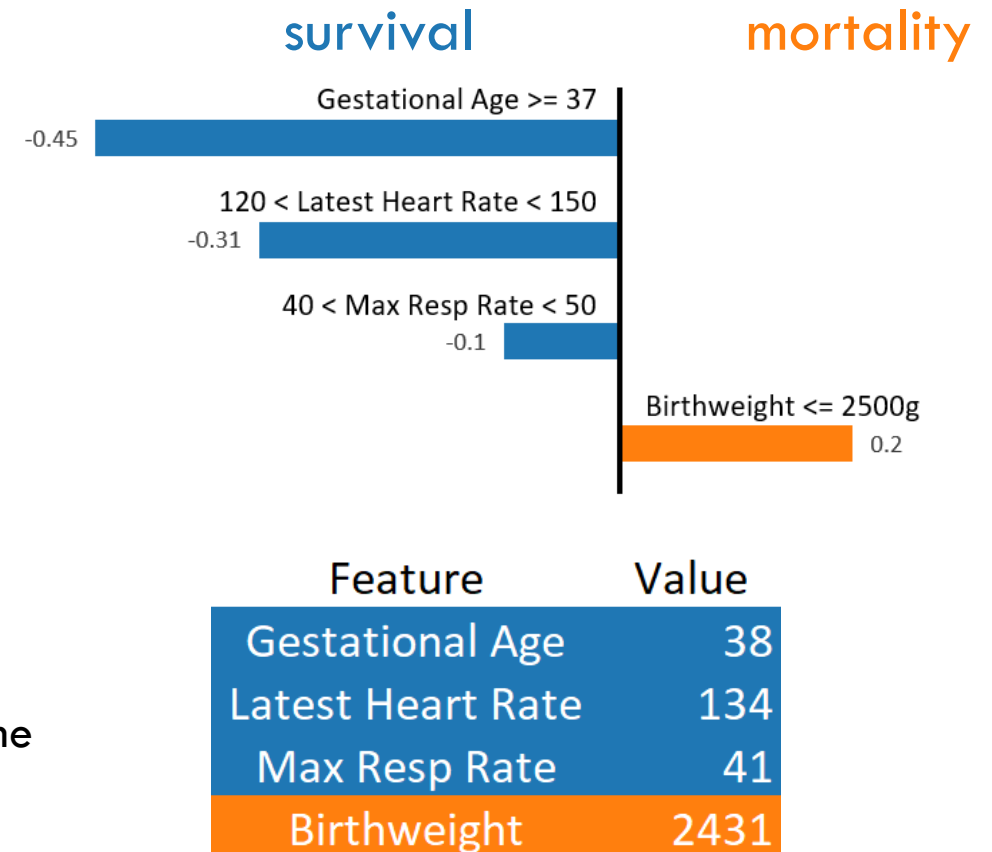
Figure Source: S. Baker, *et al.*, "Hybridized neural networks for non-invasive and continuous mortality risk assessment in neonates," *Comput. Biol. Med.*, vol. 134, p. 104521, 2021.

# Post-hoc explainability: feature importance approaches (local)

## LIME [1]

- Develops a local explanation by giving the model slight perturbations of the actual data sample to evaluate how the model's prediction will change
- A local linear model is fit to the predictions on perturbed data, with weights for each feature of this linear model then inspected to get the LIME values

[1] M. T. Ribeiro, et al., “‘Why should I trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD*, 2016, pp. 1135–1144.



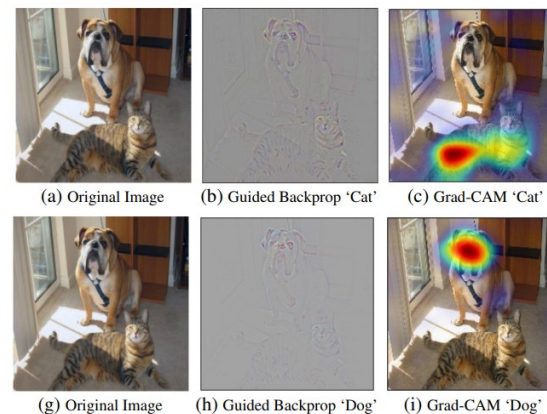
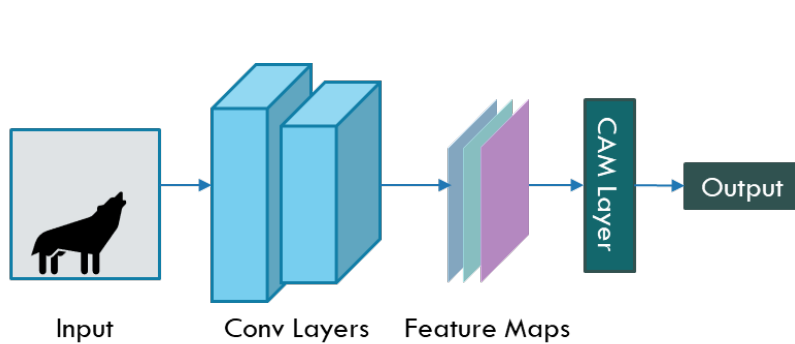
# Post-hoc explainability: explaining images and waveforms (local)

- **How can it help your research?** Analyse predictions that your model got wrong, explain your model to end users.
- Heatmapping approaches including class activation mapping (CAM) and its variants are popular in this area.
- SHAP and LIME can be applied to explaining images, but not broadly used.

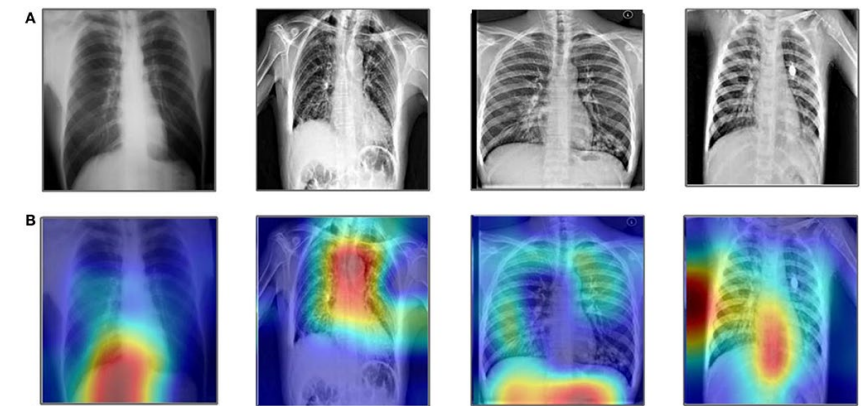
# Post-hoc explainability: explaining images and waveforms (local)

## CAM and variants

- Early version of CAM applies global average pooling on CNN feature maps before the output layer.
- Other variants take different approaches to weighting the convolutional feature maps, but most focus on the layer before reduction to an output.



Source: R. R. Selvaraju, *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE ICCV*, 2017, pp. 618–626.



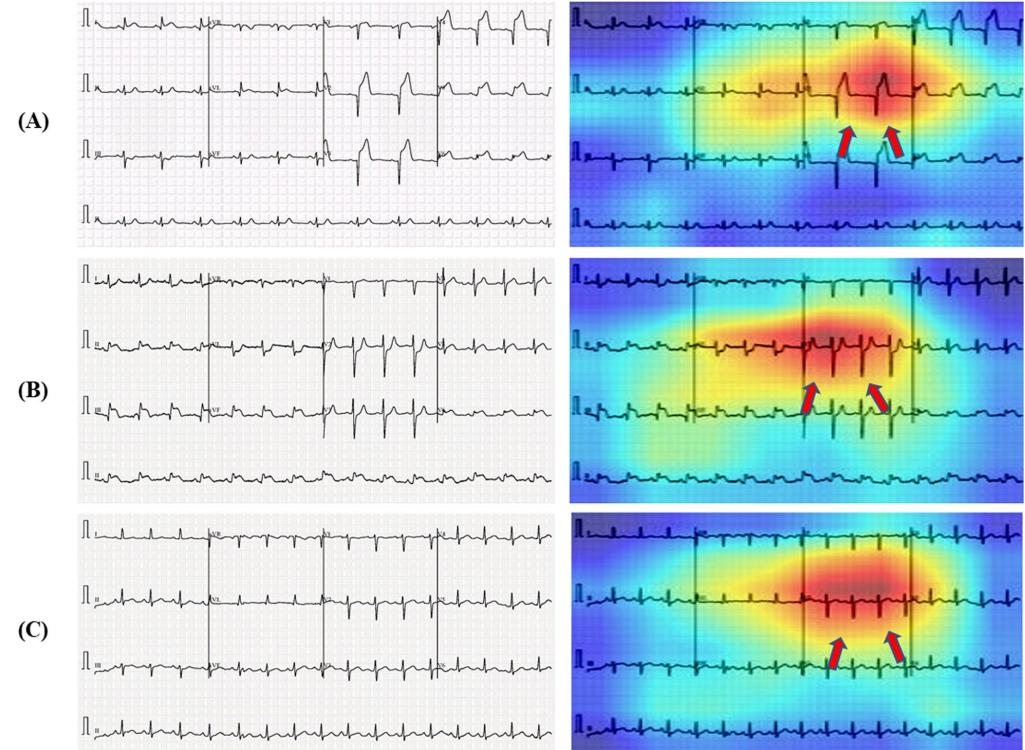
Source: A. Hamza *et al.*, "COVID-19 classification using chest X-ray images based on fusion-assisted deep Bayesian optimization and Grad-CAM visualization," *Frontiers in Public Health*, vol. 10, 2022.



# Post-hoc explainability: explaining images and waveforms (local)

## CAM and variants

- Recent papers have also applied CAM-based approaches to biomedical waveform data.

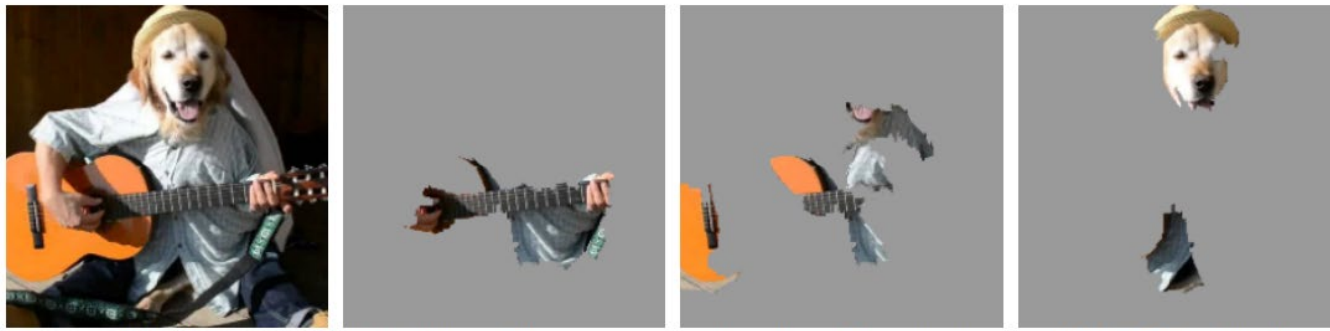


T. Rahman *et al.*, "COV-ECGNET: COVID-19 detection using ECG trace images with deep convolutional neural network," *Heal. Inf. Sci. Syst.*, vol. 10, no. 1, p. 1, 2022.

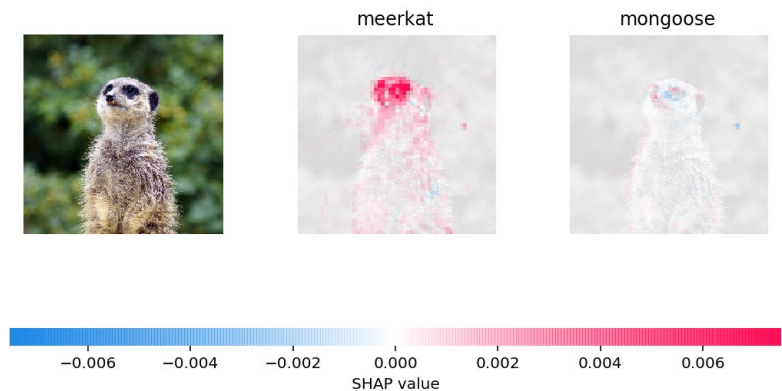
# Post-hoc explainability: explaining images and waveforms (local)

## SHAP and LIME

- Both LIME and SHAP can be used to explain how a model classified an image
- More literature seems to use LIME, likely because it's more computationally efficient
- Neither of these approaches have been broadly explored for waveform data

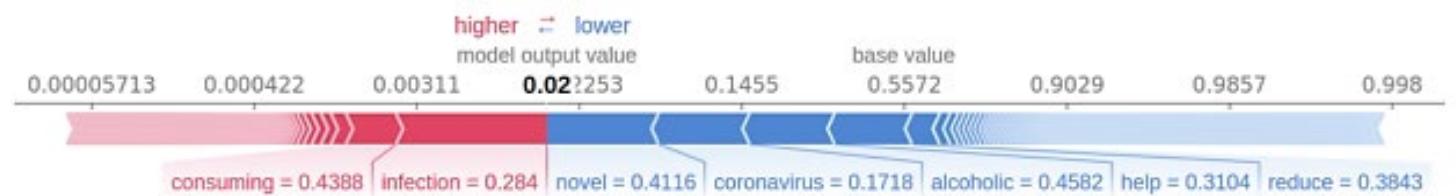


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*



# Post-hoc explainability: explaining text (local)

- Relatively few works have looked at explaining text classification.
- SHAP has been used to rank word importances for classifying misinformation [1]
- Heatmapping has been used to highlight words for sentiment analysis [2]
- Open research challenge.



Explaining classification of text statements as true or false for the statement “Consuming alcoholic beverages may help reduce the risk of infection by the novel coronavirus”. Source: [1]

[1] J. Ayoub *et al.*, “Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models,” *arXiv*, 2021.

[2] H. Chefer *et al.*, “Transformer Interpretability Beyond Attention Visualization,” in *2021 IEEE/CVF CVPR*, 2021, pp. 782–791.

# Explainability continued...

- Many more methods than the ones shown here.
- Several major open research questions:
  - How to explain generated content?
  - How do we know if explanations are truly interpretable?
  - How can we compare two different explainability approaches?





# EXPLAINABILITY FOR RESPONSIBILITY

# Explainability for responsibility

- XAI can support responsibility across all characteristics – but research is in early days

Privacy

Robustness

Transparency

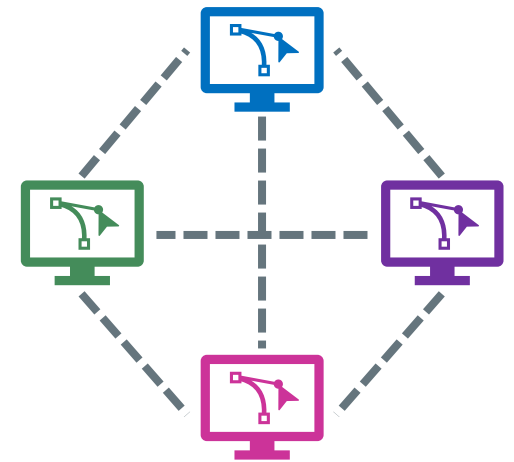
Fairness

Accountability

Safety

# Explainability for privacy

- Explainability is often considered to be a privacy risk.
- Explanations need to work alongside data protection to preserve privacy without making models harder to interpret.
- Heatmapping approaches work effectively in models that have been trained using privacy-preserving federated or swarm learning [1, 2].
- SHAP is minimally affected when differential privacy is used to anonymise training data [3].
- Explainability supports privacy indirectly – it enables privacy to be implemented without compromising other responsibility characteristics.



[1] A. Raza *et al.*, “Designing ECG monitoring healthcare system with federated transfer learning and explainable AI,” *Knowledge-Based Syst.*, vol. 236, p. 107763, 2022.

[2] O. L. Saldanha *et al.*, “Swarm learning for decentralized artificial intelligence in cancer histopathology,” *Nat. Med.*, vol. 28, no. 6, pp. 1232–1239, 2022.

[3] A. Bozorgpanah *et al.*, “Privacy and Explainability: The Effects of Data Protection on Shapley Values,” *Technologies*, vol. 10, no. 6, 2022.

# Explainability for robustness

- Explainability supports testing for consistency and resilience against attacks.
- Inspecting individual explanations can reveal issues with consistency, and metrics have been proposed which use explainability to quantify robustness [1].
- Some research has suggested that robust models have more interpretable salience maps [2].

Label	Sample	NR, simp	NR, SmG + thresh	R, simp	R, SmG + thresh	Label	Sample	NR, simp	NR, SmG + thresh	R, simp	R, SmG + thresh
eight						dog					
three						bird					
seven						dog					
one						air-plane					
four						truck					

Source: [3]

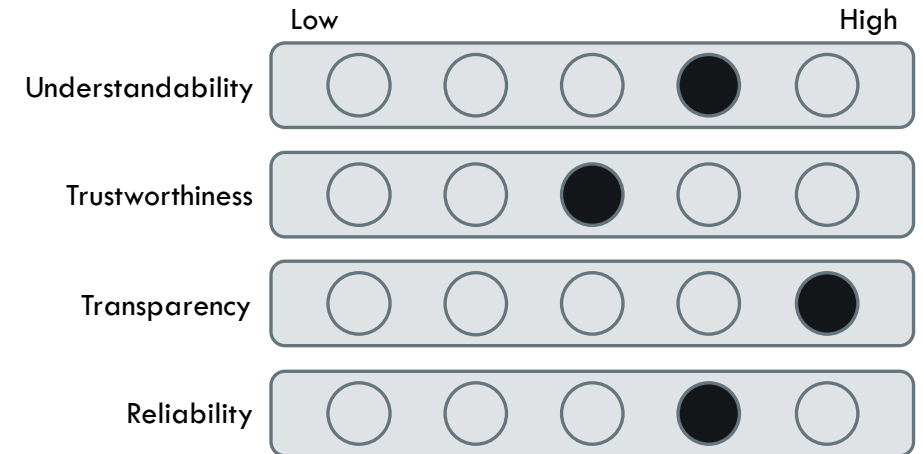
[1] S. Sharma *et al.*, "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models," in AAAI/ACM AIES '20, 2020, pp. 166–172

[2] A. Noack *et al.*, "An Empirical Study on the Relation Between Network Interpretability and Adversarial Robustness," *SN Comput. Sci.*, vol. 2, no. 1, p. 32, 2021.



# Explainability for transparency

- Transparency is the original goal of XAI.
- Some studies have assessed XAI models are truly transparent, largely using end-user surveys [1-3].
- Explanations generally improved perceived understandability and transparency.
- Relatively few works have looked to quantify the genuine transparency of XAI approaches.



Example of survey approach

[1] E. Khodabandehloo *et al.*, “HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline,” *Futur. Gener. Comput. Syst.*, vol. 116, pp. 168–189, 2021.

[2] S. Singla *et al.*, “Explaining the black-box smoothly—A counterfactual approach,” *Med. Image Anal.*, vol. 84, p. 102721, 2023.

[3] M. Naiseh *et al.*, “How the different explanation classes impact trust calibration: The case of clinical decision support systems,” *Int. J. Hum. Comput. Stud.*, vol. 169, p. 102941, 2023.

# Explainability for fairness

- Explanations can reveal biases at local and global levels, allowing these to be rectified.
- LIME has been used to revealed biases in justice system AI models [1].
- Rule-based explainability has been proposed as a tool for ‘fairness auditing’ [2, 3].
- Some research suggests that some post-hoc methods exhibit unfairness [4] – meaning there is a research opportunity to improve on this.

[1] M. Miron *et al.*, “Evaluating causes of algorithmic bias in juvenile criminal recidivism,” *Artif. Intell. Law*, vol. 29, no. 2, pp. 111–147, 2021.

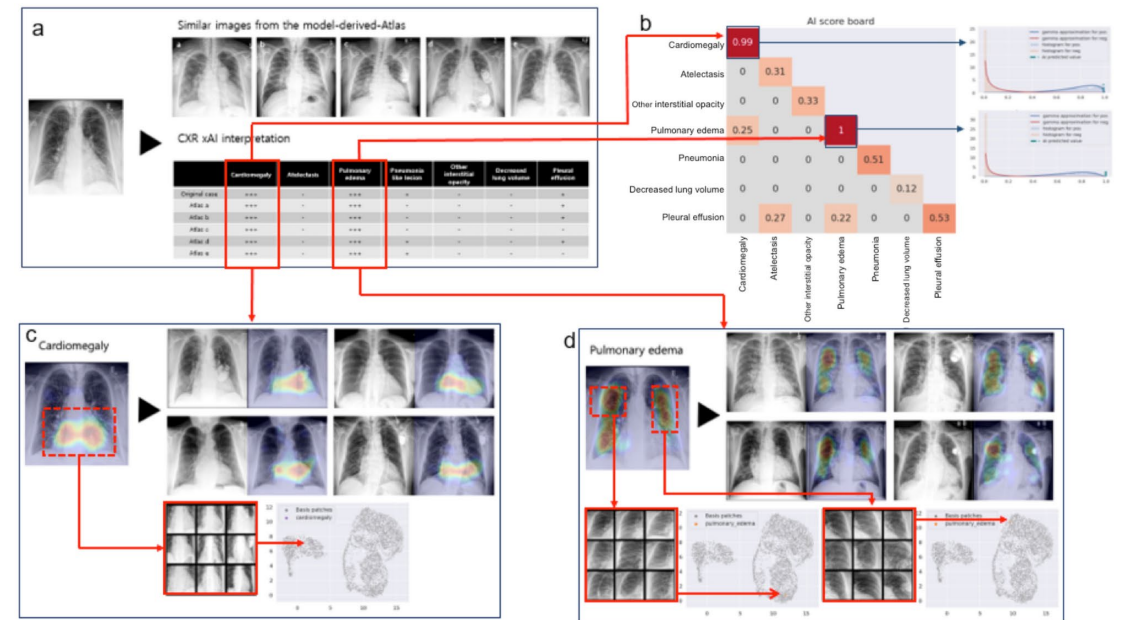
[2] C. Panigutti *et al.*, “FairLens: Auditing black-box clinical decision support systems,” *Inf. Process. Manag.*, vol. 58, no. 5, p. 102657, 2021.

[3] C. Panigutti *et al.*, “Doctor XAI: an ontology-based approach to black-box sequential data classification explanations,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 629–639.

[4] J. Dai *et al.*, “Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 203–214.

# Explainability for accountability

- Explainability can help end-users to hold AI accountable for its decisions
- GradCAM has been used to support auditability in x-ray classification systems [1].
- Explainable-by-design logistic regression models have been used to explain decisions to human end-users, who can then provide their feedback on the fairness of a decision [2].



Source: [1]

[1] J. Chung *et al.*, "Prediction of oxygen requirement in patients with COVID-19 using a pre-trained chest radiograph xAI model: efficient development of auditable risk prediction models via a fine-tuning approach," *Sci. Rep.*, vol. 12, no. 1, p. 21164, 2022.

[2] Y. Nakao *et al.*, "Toward Involving End-users in Interactive Human-in-the-loop AI Fairness," *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 3, pp. 1–30, 2022.

# Explainability for safety

- Explainability supports safety in critical decision making, and in detecting attacks.
- Explanations have been shown to improve patient safety by supporting doctors in understanding risk levels through counterfactuals and feature importances [1].
- LIME and SHAP have been shown to assist in detection and resolution of IoT network attacks [2].

[1] Y. Jia *et al.*, “The Role of Explainability in Assuring Safety of Machine Learning in Healthcare,” *IEEE Trans. Emerg. Top. Comput.*, vol. 10, no. 4, pp. 1746–1760, 2022.

[2] Z. A. E. Houda *et al.*, “‘Why Should I Trust Your IDS?’: An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks,” *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1164–1176, 2022.

# Using XAI to make responsible AIs

- Early works have shown that explainability supports responsibility in many ways
- ... but the field is still in its infancy.
- **There is significant research opportunity in this space.**

# Using XAI to make responsible AIs

## Some thoughts on next research steps

- Responsibility can be improved in any AI work by incorporating explanations for improved transparency and assessment of robustness and fairness.
- Metrics are needed for quantifying whether explanations are genuinely interpretable - good explainability is essential for responsibility.
- Metrics and frameworks are needed for quantifying responsibility across all metrics – explainability has shown promise in developing these.
- Generative AIs are a problem area – need explanation methods that support responsible GenAI development and use.
- Human-in-the-loop feedback is an interesting but challenging direction for responsible AI



# RESPONSIBILITY IN MY RESEARCH

# My first steps towards responsible AI

- Used SHAP to analyse predictions in recent papers.
  - This also helped to reveal problems in early iterations of my model!
- Included a simple confidence score to help clinicians assess my model's certainty in a decision.
- Wrote a review paper that is currently under review.
- Next focus: quantifying explainability towards assessing which XAI methods are most useful in responsible AI.

$$\text{Confidence Percentage} = \frac{|0.5 - \text{output}|}{0.5} \times 100$$

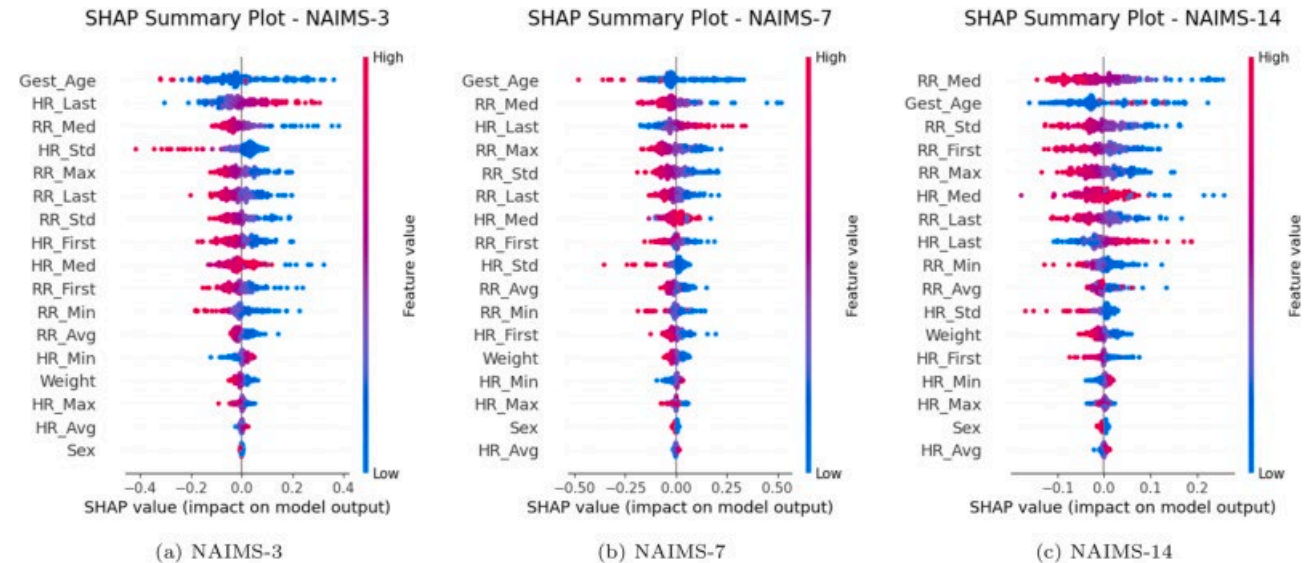
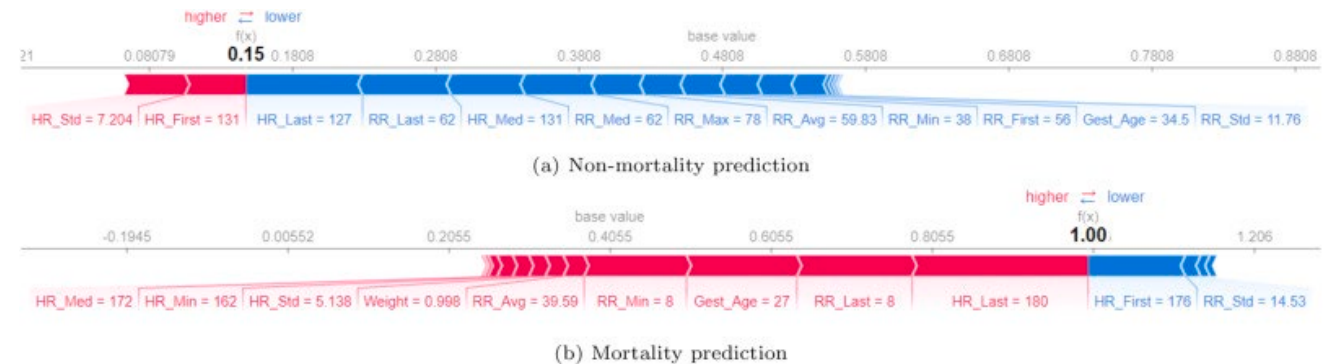
Explanation at its simplest.

Source: S. Baker *et al.*, "Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach," *Sci. Rep.*, vol. 10, no. 1, p. 21282, 2020.



# My first steps towards responsible AI

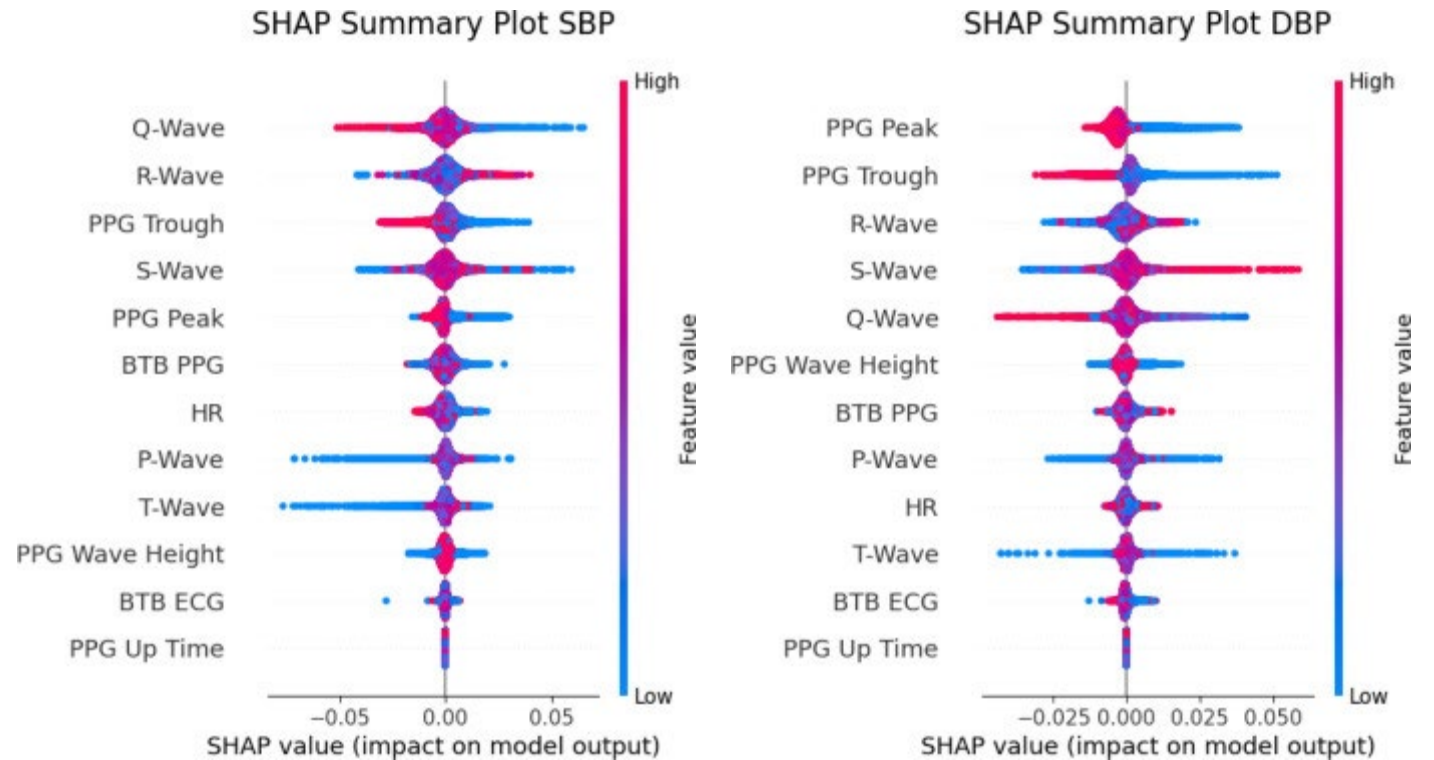
- SHAP analysis is my most-used tool – I like the global and local contexts available.



Source: S. Baker *et al.*, “Hybridized neural networks for non-invasive and continuous mortality risk assessment in neonates,” *Comput. Biol. Med.*, vol. 134, p. 104521, 2021.

# My first steps towards responsible AI

- SHAP analysis is my most-used tool – I like the global and local contexts available.



Source: S. Baker *et al.*, "A computationally efficient CNN-LSTM neural network for estimation of blood pressure from features of electrocardiogram and photoplethysmogram waveforms," *Knowledge-Based Syst.*, vol. 250, p. 109151, Aug. 2022.

# Why this field?

- The AI explosion has meant lots of tools exist to make AI development easier.
- Hard to stand out by just applying AI to a new problem.
- Explanation helps to prove that your AI is good, while also supporting responsibility.
- Responsibility is hugely important to industry and governments – huge research opportunities in works that improve aspects of responsibility.



# QUESTIONS OR COMMENTS?



Dr Stephanie Baker  
James Cook University

E: [stephanie.baker@jcu.edu.au](mailto:stephanie.baker@jcu.edu.au)  
P: +61 7 4232 1561

