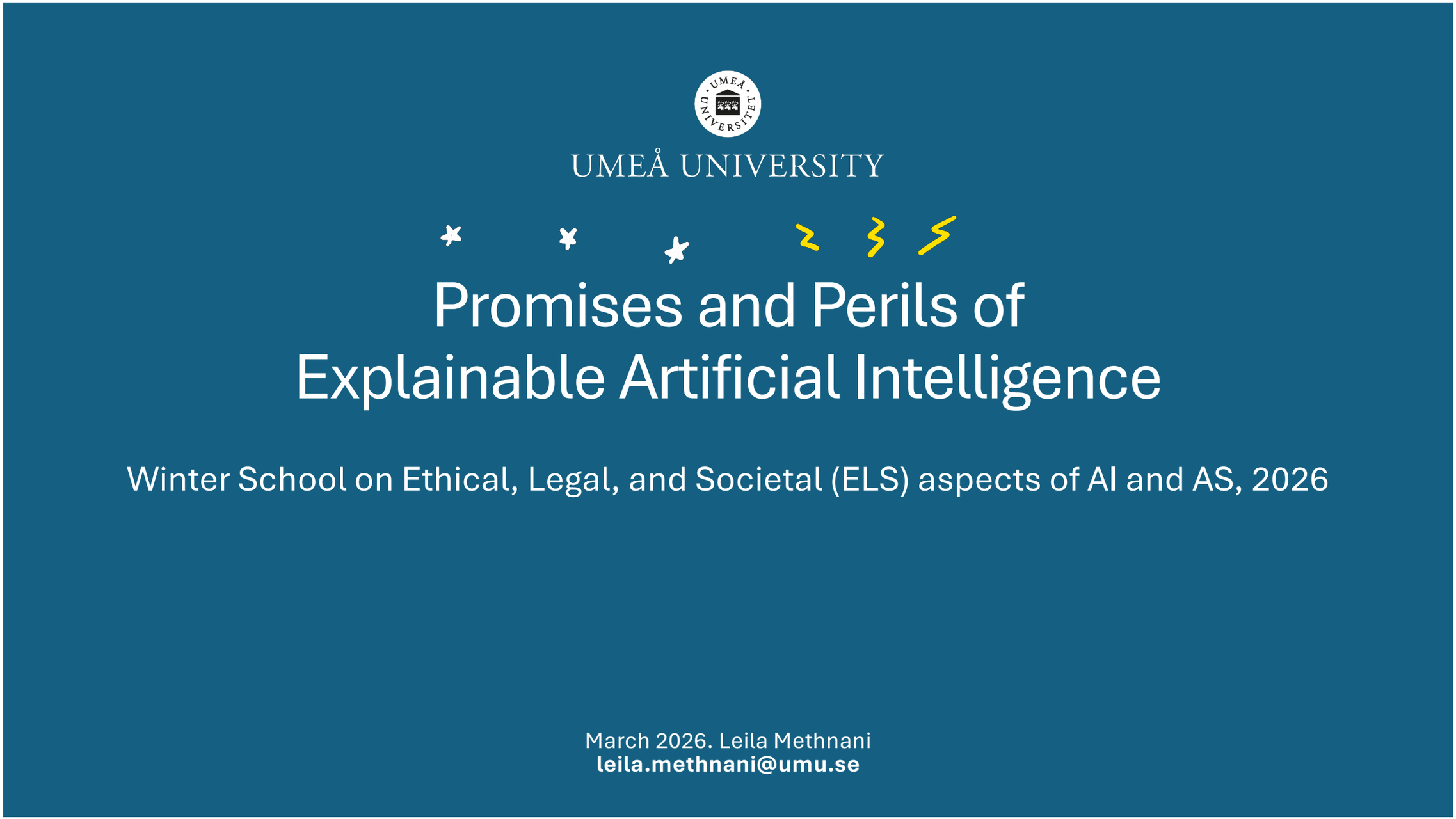



# Promises and Perils of Explainable Artificial Intelligence

Winter School on Ethical, Legal, and Societal (ELS) aspects of AI and AS, 2026

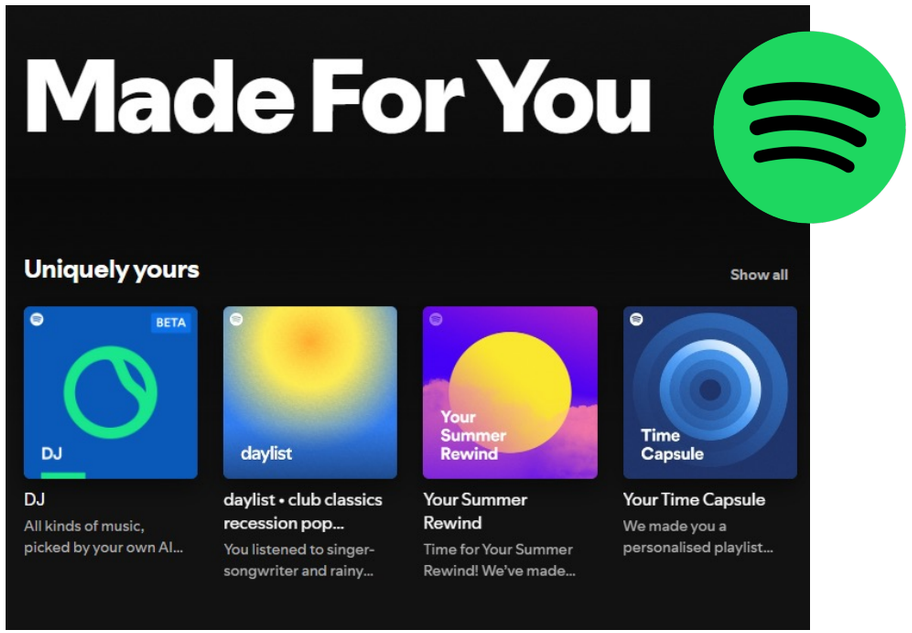
March 2026. Leila Methnani
**leila.methnani@umu.se**

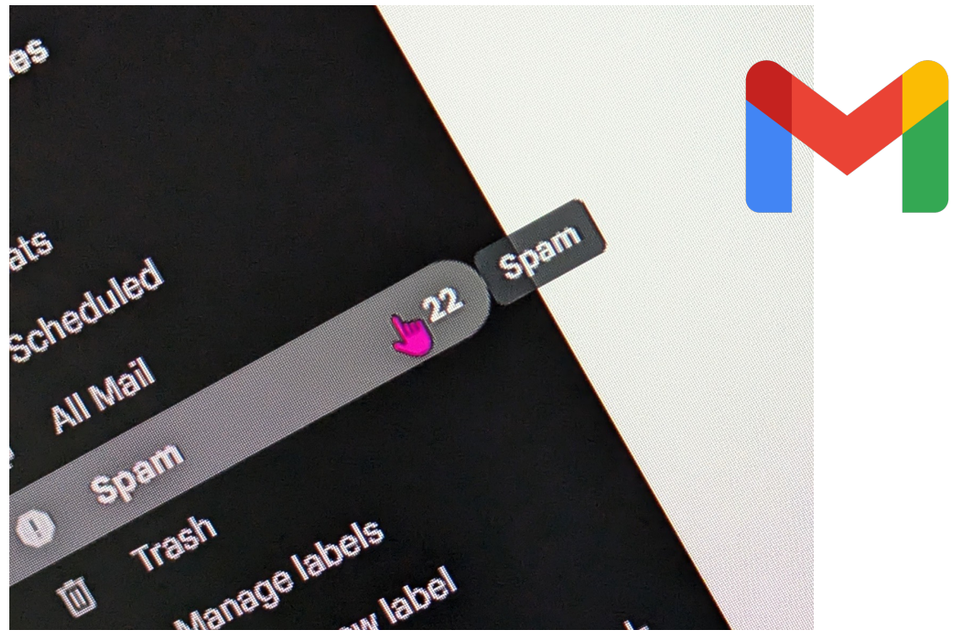
# What we will cover today

- GDPR requires a “right to explanation” + AI Act highlights interpretability.
- What is an “explanation”? and eXplainable AI (XAI)?
- Look at what XAI aims to achieve for diverse stakeholder needs.
- Consider numerous XAI approaches; selection of technique will depend on target *explainee* needs.
- Highlight that XAI is a double-edged sword! It should also be handled with care.

# Automated Decision-making

- Used in a range of daily contexts we might not even think about:

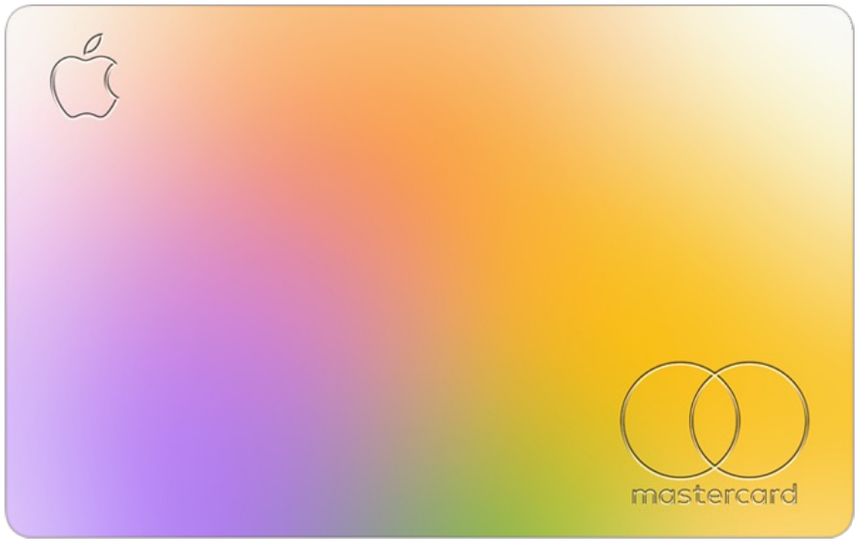



Screenshot (left) and photo (right) taken by LM Aug 2024.

# Automated Decision-making

- Used across industries like finance and healthcare:



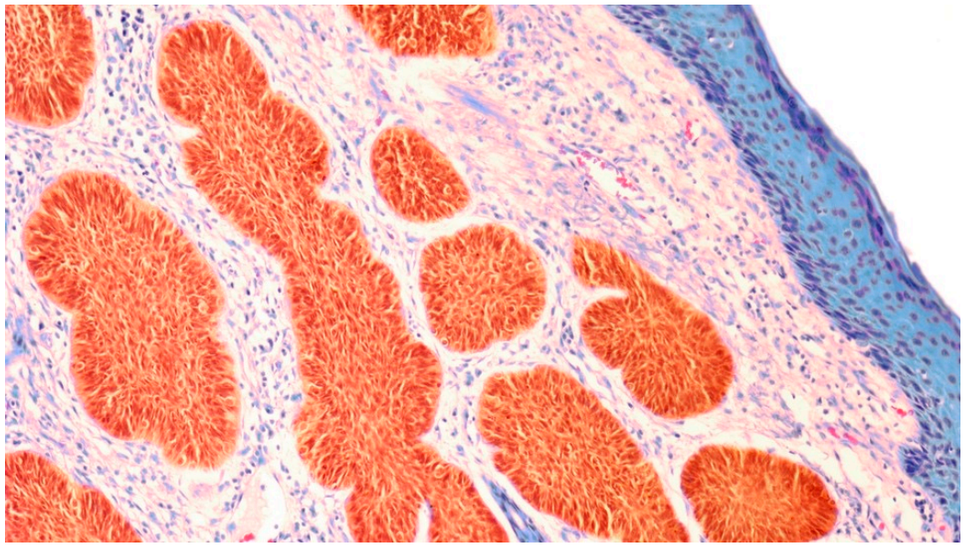



Photo sources: support.apple.com and Steve Gschmeissner / Getty

# “High-risk” makes the difference

- Frustrating user experience doesn’t have same impact on quality of life:

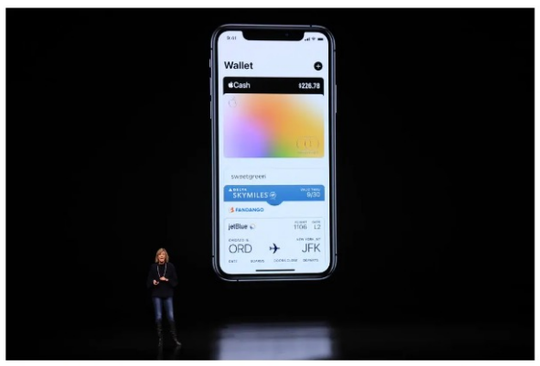
The New York Times

## *Apple Card Investigated After Gender Discrimination Complaints*

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

By Neil Vigdor

Nov. 10, 2019



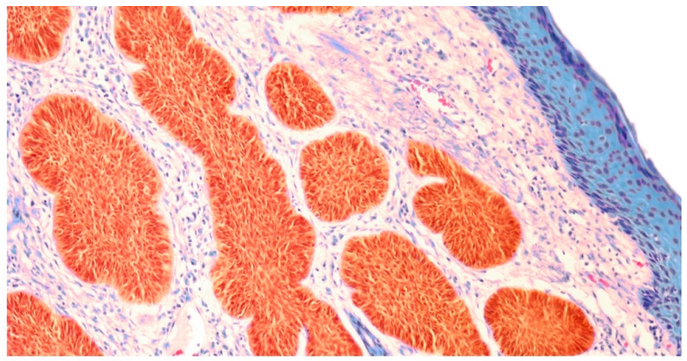

Jennifer Bailey, vice president of Apple Pay. Regulators are investigating Apple Card’s algorithm, which is used to determine applicants’ creditworthiness. Jim Wilson/The New York Times

Popular Latest Newsletters *The Atlantic* Sign In

## AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

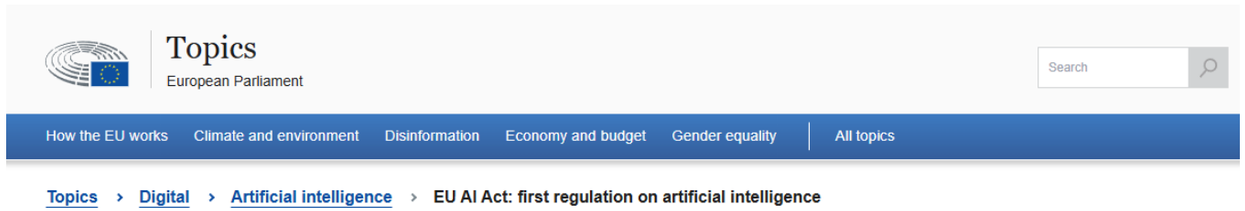
By Angela Lashbrook



Steve Gschmeissner / Getty

Articles accessed and screenshots taken Aug 2024:
nytimes.com; The Atlantic.

# The EU AI Act is intended to protect you




## EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023
Last updated: 18-06-2024 - 16:29
6 min read

**Table of contents**



Page accessed and screenshot taken Aug 2024: EU AI Act: first regulation on artificial intelligence | Topics | European Parliament (europa.eu).

# Still many questions around Article 14

- **“The AI system should be provided in a way that allows the overseer to understand its capabilities and limitations, detect and address issues, avoid over-reliance on the system, interpret its output, decide not to use it, or stop its operation.”**
- Explainability and interpretability are indeed critical for such systems and mechanisms, even (or especially) with the human present.
- Explainable AI tools have a long way to go…

EU AI Act: first regulation on artificial intelligence | Topics | European Parliament (europa.eu).

# Opaque Models

- Powerful and useful, but these models are opaque: not interpretable to humans.
- Daily human decision-making informed by these models.
  - *False positive diagnosis – Patient suffers through difficult treatment*.
  - *False negative diagnosis – Patient untreated for prolonged period*.
- XAI addresses this challenge for stakeholders.

# What is Explainable AI?

- XAI is a field of research.
- Develop methods and techniques that explain AI decision-making.
- Aims to make opaque AI systems understandable to humans
  - How? By offering explanations.

Page accessed and screenshot taken Aug 2024: EU AI Act: first regulation on artificial intelligence | Topics | European Parliament (europa.eu).

# Explanation:

*“an interface between humans and a decision maker that is … both an accurate proxy of the decision maker and comprehensible to humans.*

— R. Guidotti et al.

# XAI for who?

- Receiver of explanation influences explainability.
- Developer vs. User vs. Investor vs. Governing bodies, etc.

**“I need to debug the system. Why is it behaving this way?”**

# XAI for who?

- Receiver of explanation influences explainability.
- Developer vs. User vs. Investor vs. Governing bodies, etc.

**“How do I know this is a trustworthy diagnosis?”**

# XAI for who?

- Receiver of explanation influences explainability.
- Developer vs. User vs. Investor vs. Governing bodies, etc.

# XAI for who?

- Receiver of explanation influences explainability.
- Developer vs. User vs. Investor vs. Governing bodies, etc.

**“Is this system fair?”**

# XAI for who?

- Receiver of explanation influences explainability.
- Developer vs. User vs. Investor vs. Governing bodies, etc.

**“Will the industry even adopt this system if they cannot understand it’s output?”**

# Why else do we need XAI?

- Support AI uptake in industry
- Calibrate trust
- Knowledge acquisition
- Fairness
- Accessibility
- Interactivity
- Amongst others …

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion* 58 (2020): 82-115.

# Interpretability

- Interpretability != explainability?

*“One of the issues that hinders the establishment of common grounds is the interchangeable misuse of interpretability and explainability in the literature.”*

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion* 58 (2020): 82-115.

# Interpretability

- The literature cannot seem to agree on the difference between explainability and interpretability.

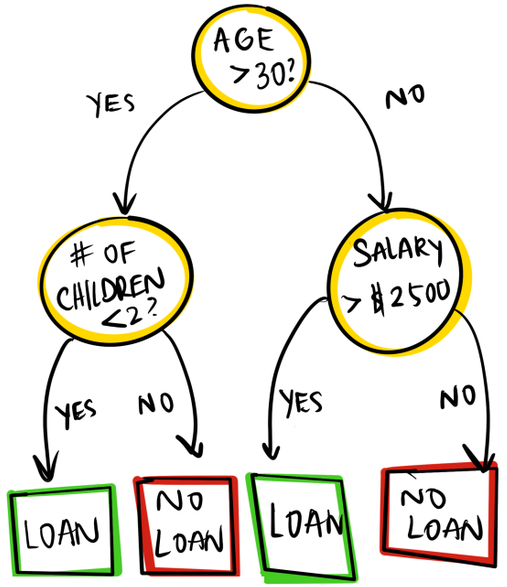
*“A model is more interpretable than another model if its decisions are easier for a human to understand than the other model’s decisions.”*

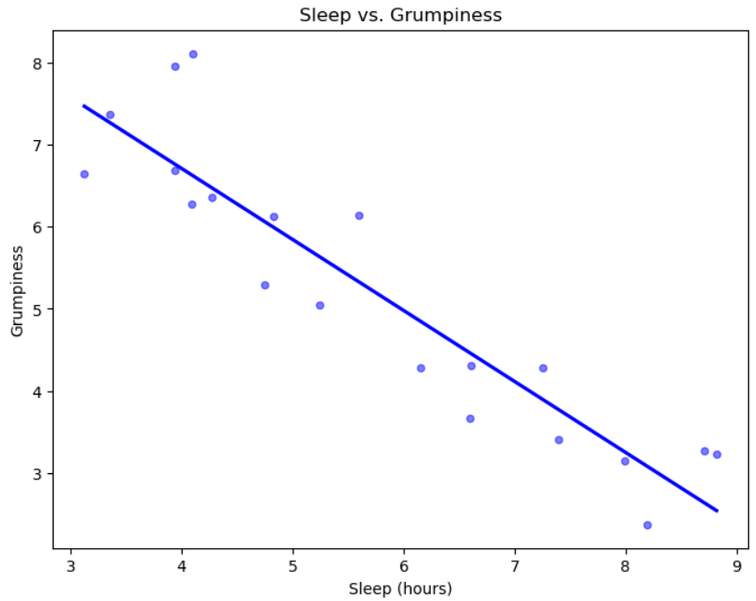
Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3rd ed., 2025. ISBN: 978-3-911578-03-5. Available at: https://christophm.github.io/interpretable-ml-book.

# Interpretable models

- An interpretable model is one that is easily human understandable, e.g.:
  - decision trees
  - linear models

- Contrasted with post-hoc explainability.



# Is there an interpretability vs accuracy trade-off?

*“**It is a myth that there is necessarily a trade-off between accuracy and interpretability…**When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing.”*

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1.5 (2019): 206-215.

# Is there an interpretability vs accuracy trade-off?

*“It is important not to assume that one needs to make a sacrifice in accuracy in order to gain interpretability.”*

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1.5 (2019): 206-215.

# Is there an interpretability vs accuracy trade-off?

- “Sparser” or “simpler” models are a subset of interpretable models, but they are not the only option.

- Many are working towards building interpretable models e.g.
  - Generalized Additive Models (GAMs) -> extension of linear models
  - Concept Bottleneck Models

Koh, Pang Wei, et al. "Concept bottleneck models." *International conference on machine learning*. PMLR, 2020.

# Is there an interpretability vs accuracy trade-off?

*"We show that bottleneck models are comparable to standard end-to-end models while also attaining high concept accuracies."*

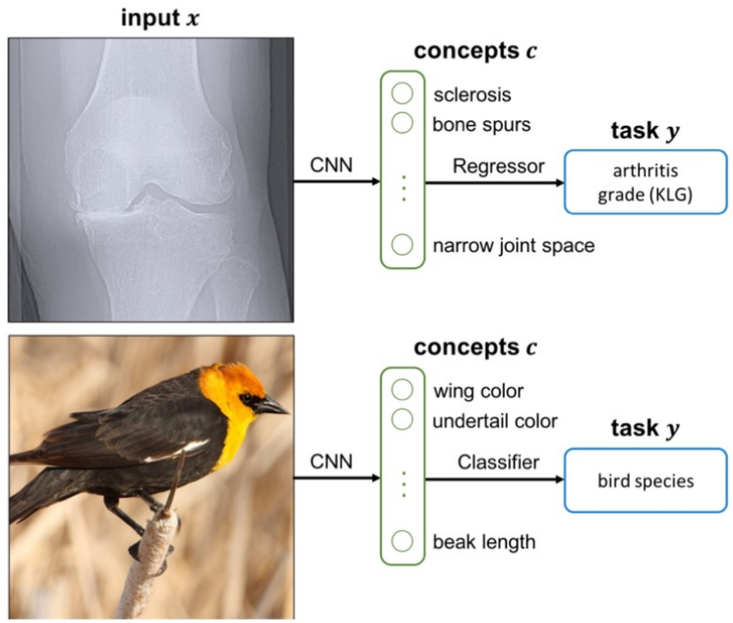



*Figure 1.* We study concept bottleneck models that first predict an intermediate set of human-specified concepts $c$, then use $c$ to predict the final output $y$. We illustrate the two applications we consider: knee x-ray grading and bird identification.

Koh, Pang Wei, et al. "Concept bottleneck models." *International conference on machine learning*. PMLR, 2020.

# Ask yourself …

- Is an AI solution even warranted in this situation?
- Can we fit an interpretable model that performs well?
- If no, what explainable techniques are appropriate to support post-hoc explainability of these complex models?

# Classification of XAI methods: scope

- Global explanations:
  - *How does the model use the concept of stripes to classify images?*
- Local explanations
  - *Why was Sarah denied a loan?*
- Cohort explanations
  - *How does model output differ between people who live South of the river vs. North of the river?*

# Classification of XAI methods: where they can be applied

- Model specific
  - Usually exploit architectural characteristics in explanation generation.

- Model agnostic
  - Does not care about architecture; treated as a “black box”

# Classification of XAI methods: the shape of the explanation or how it is generated

- Feature-based
  - What feature is important for this model outcome?
- Concept-based
  - Higher level, human-understandable representations.
- Example-based
  - Extract or generate examples representative of the data.
- …

# Local Interpretable Model-Agnostic Explanations (LIME)

- Local: a single instance is explained faithfully (vs. global explainability).
- Model agnostic: can be applied to any model (vs. model-specific).
- Fit a “surrogate” interpretable model in the local perturbed neighbourhood of a single instance and use the new model as an explanation.

Image source: https://arize.com/glossary/local-interpretable-model-agnostic-explanations-lime


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016./

# LIME is “locally faithful” but unstable

- “LIME’s goal is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier.”
- “Faithful”: explanation produces the same prediction outcome as the opaque model.
- Different explanations for similar instances (Alvarez-Melis and Jaakkola 2018)

Image source: https://arize.com/glossary/local-interpretable-model-agnostic-explanations-lime


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016./

# LIME can be manipulated

- Adversarial attack on post-hoc perturbation-based explanations (Slack et al. 2020)
- Instability and ease of manipulation contribute to decreased trust in the explanation method.
- Strength? intuitive.

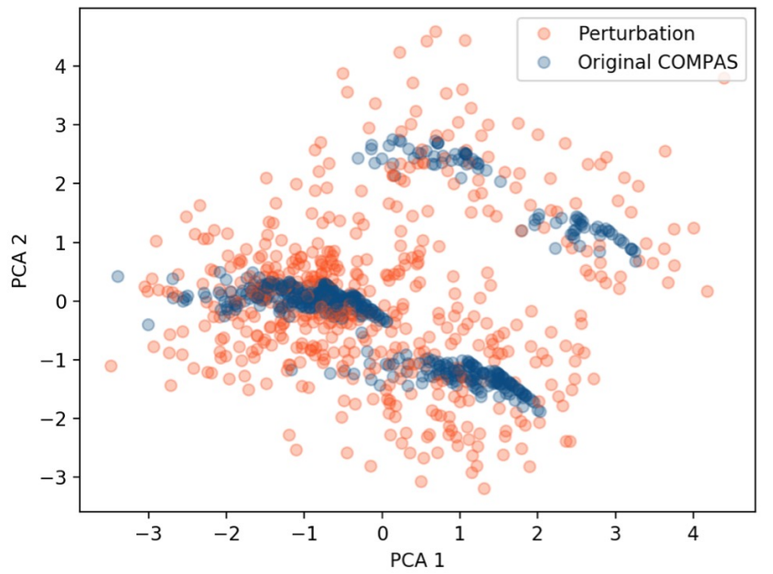



**Figure 1: PCA applied to the COMPAS dataset (blue) as well as its LIME style perturbations (red). Even in this low-dimensional space, we can see that data points generated via perturbations are distributed very differently from instances in the COMPAS data. In this paper, we exploit this difference to craft adversarial classifiers.**

Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.

# Example-based explanations


Comparative


Normative

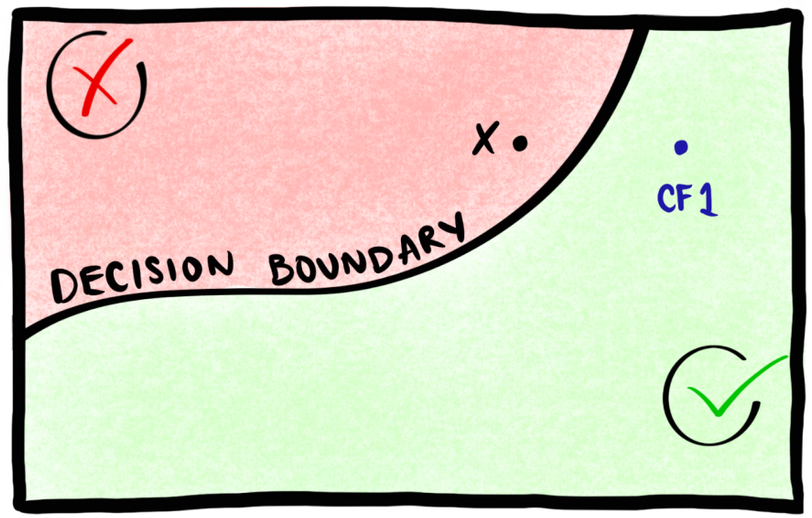
- Cai et al. "The effects of example-based explanations in a machine learning interface.“ (2019)

# Counterfactual explanations

- Aim to answer `why not P instead of Q'.

- Perform minimal feature changes until an alternative prediction is made.

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |

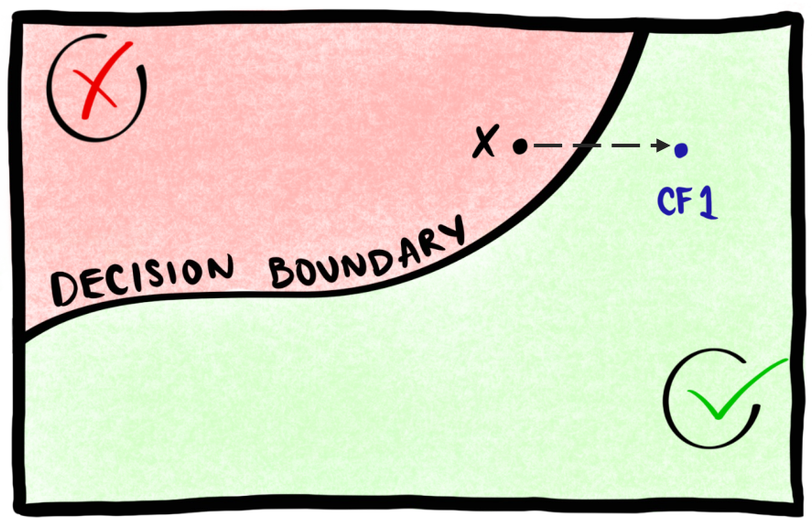


# Counterfactual explanations

- Aim to answer `why not P instead of Q'.

- Perform minimal feature changes until an alternative prediction is made.

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |

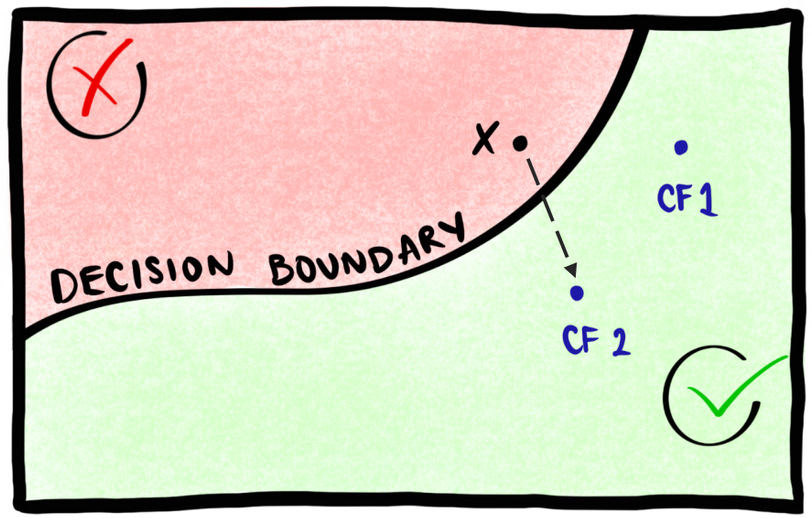


# Counterfactual explanations

- Aim to answer `why not P instead of Q'.

| ... | Education | Years Experience | Ethnicity | ... |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |

- Perform minimal feature changes until an alternative prediction is made.




- Wachter et al. "Counterfactual explanations without opening the black box" 2017.

# Rashomon Effect

- Many, contradicting, accounts of the same event.

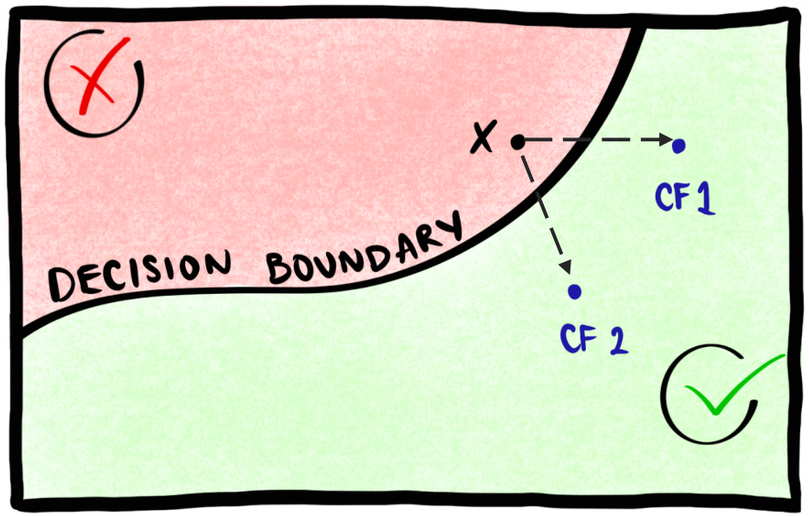

- There may be several *valid* changes to the input that alter the outcome, which one is *best?*

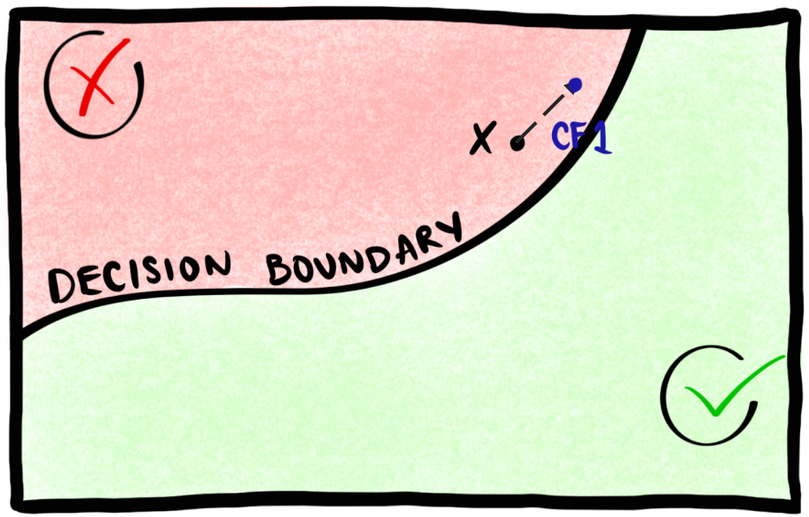
Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3rd ed., 2025. ISBN: 978-3-911578-03-5. Available at: https://christophm.github.io/interpretable-ml-book.

# Properties of counterfactual explanations

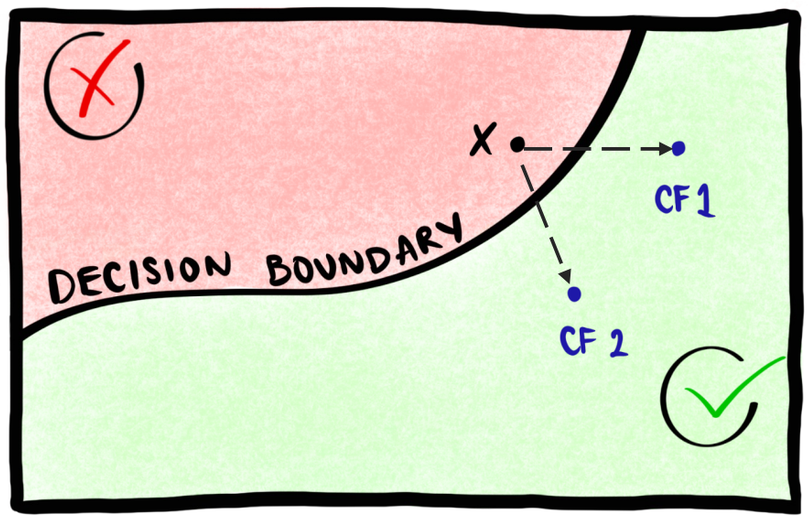
- **Validity:** actually flips the outcome
- **Sparsity:** minimal feature changes
- **Feasibility/ Plausibility:** possible or practical to do
- **Actionability:** adheres to constraints on what can actually be changed
- **Diversity:** generates many different valid counterfactuals
- **Proximity:** remains similar to the original input sample
- … and more…

# Valid vs invalid



# Sparse vs … less sparse




| … | Education | Years Experience | … | … |
|---|---|---|---|---|
| … | … | … | … | … |

| … | Education | Years Experience | … | … |
|---|---|---|---|---|
| … | … | … | … | … |

# Diverse vs not diverse

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | 5 | ... | ... |

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | 6 | ... | ... |

| ... | Education | Years Experience | ... | ... |
|---|---|---|---|---|
| ... | ... | 7 | ... | ... |

- Diversity is to maximise information to support explainability and choice.

# Lack of psychological grounding

- M. Keane et al discuss some key deficits of counterfactuals.
- What’s plausible?
- How sparse is sparse enough?
- “The neglect of user studies is the “original sin” of XAI research.”

Keane, Mark T., et al. "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques." *arXiv preprint arXiv:2103.01035* (2021).

# Example-based explanations

Or evaluate with humans! Which is encouraged…


Comparative

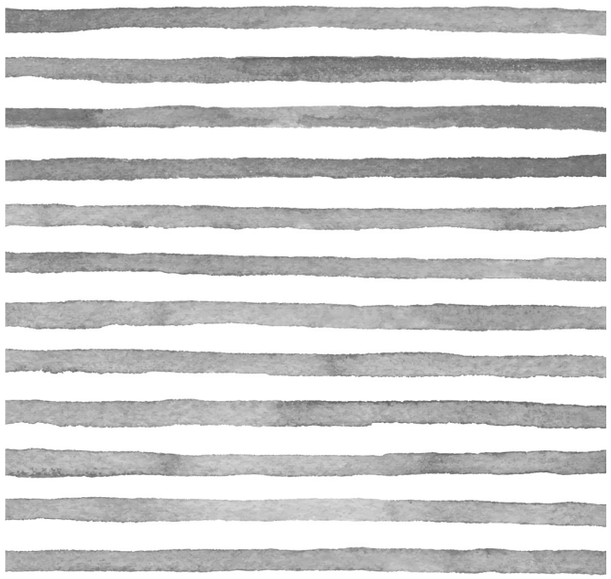

Normative

Cai et al. "The effects of example-based explanations in a machine learning interface.“ (2019)

# Mental Model:

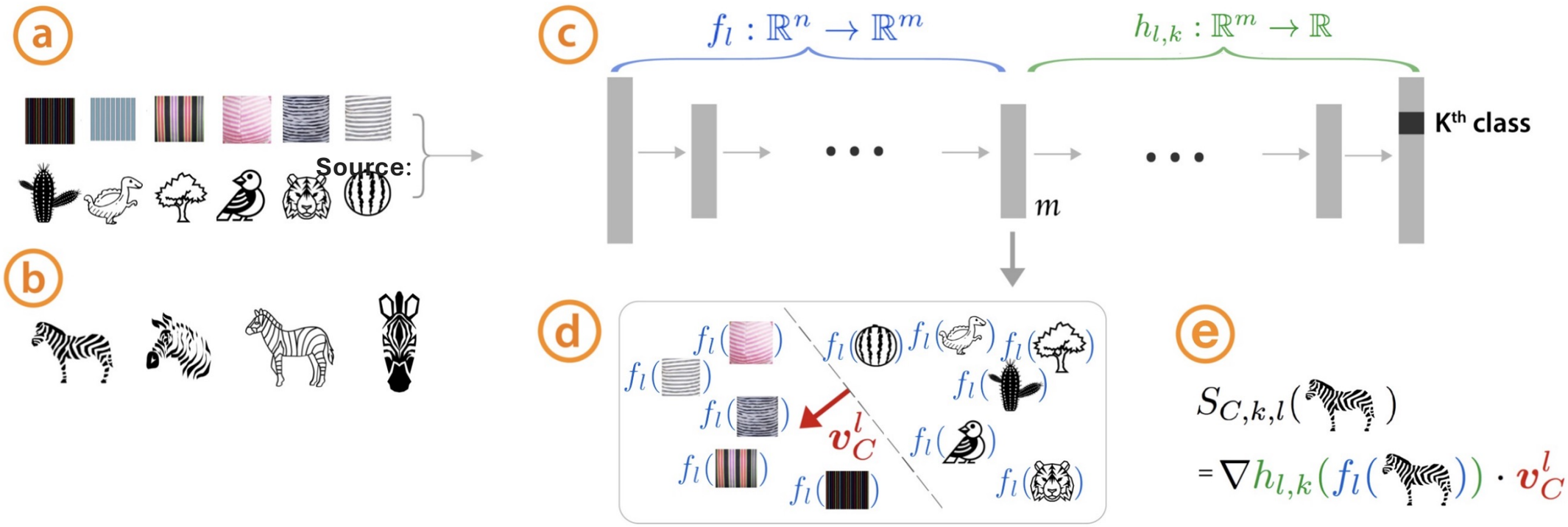
*“any internal representation of the relations between a set of elements … [such as] expectations regarding use and consequences … used to guide the individual’s interactions with the system or product in question.”*

—American Psychological Association.



***How much does the concept of "stripes" contributes to the model's prediction of Zebra?***

# Concept-based explanations

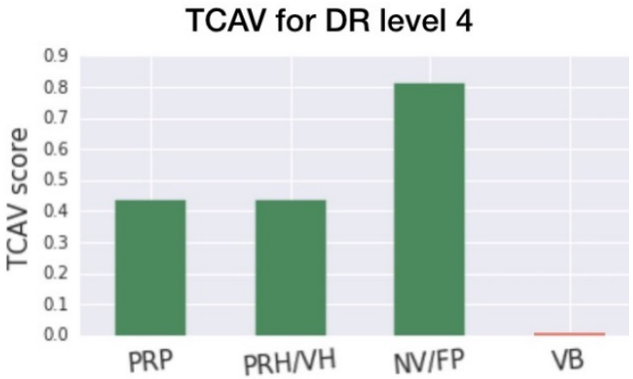



# Concept-based explanations

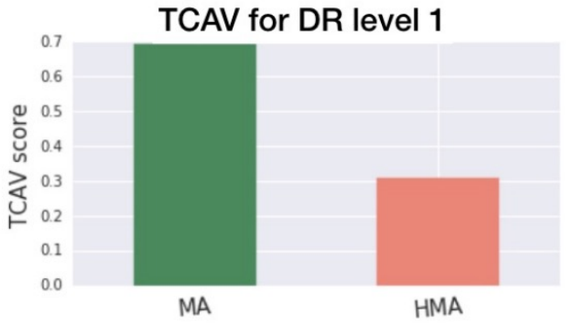
Kim et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." (2018)

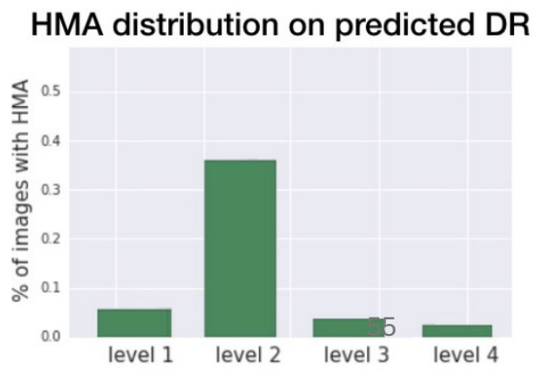
# Concepts in medicine

- diabetic retinopathy (DR) from none (0) to proliferative (4)
- Concepts such as
  - Microaneurysms (MA)
  - Pan-retinal laser scars (PRP)
- Different concepts more prominent at different DR levels
- “Given this, the doctor said they would like to tell the model to de-emphasize the importance of HMA for level 1.”






Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).” (2018)

# Lessons from the social sciences

- Contrastive
- Short
- Causal
- **Social**

## Explanation:

*“an interface between humans and a decision maker that is … both an accurate proxy of the decision maker and comprehensible to humans.*

**— R. Guidotti et al.**

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

# How we interact

- Mental models of systems are impressionable and continue to be moulded through interaction.
- Explainability changes the way we interact with the system.
- Can determine our *control* over decision-making.

# XAI’s design problem

*“While XAI and transparency mechanisms are widely proposed as solutions to improve human–AI interaction, the empirical findings thus far presented suggest that current approaches may unintentionally amplify [Automation Bias] by fostering misplaced trust.”*

Romeo, Giuseppe, and Daniela Conti. "Exploring automation bias in human–AI collaboration: a review and implications for explainable AI." *AI & SOCIETY* 41.1 (2026): 259-278.

# Automation Bias:

“a cognitive phenomenon where humans display an overreliance on automated systems, favoring automated recommendations over their own judgment, even when contradictory and more accurate information is available.”

— G. Romeo & D. Conti

Romeo, Giuseppe, and Daniela Conti. "Exploring automation bias in human–AI collaboration: a review and implications for explainable AI." *AI & SOCIETY* 41.1 (2026): 259-278.

# XAI’s design problem

*“We propose explanation design strategies that actively promote critical engagement and independent verification. ”*

*Romeo, Giuseppe, and Daniela Conti. "Exploring automation bias in human–AI collaboration: a review and implications for explainable AI." AI & SOCIETY 41.1 (2026): 259-278.*

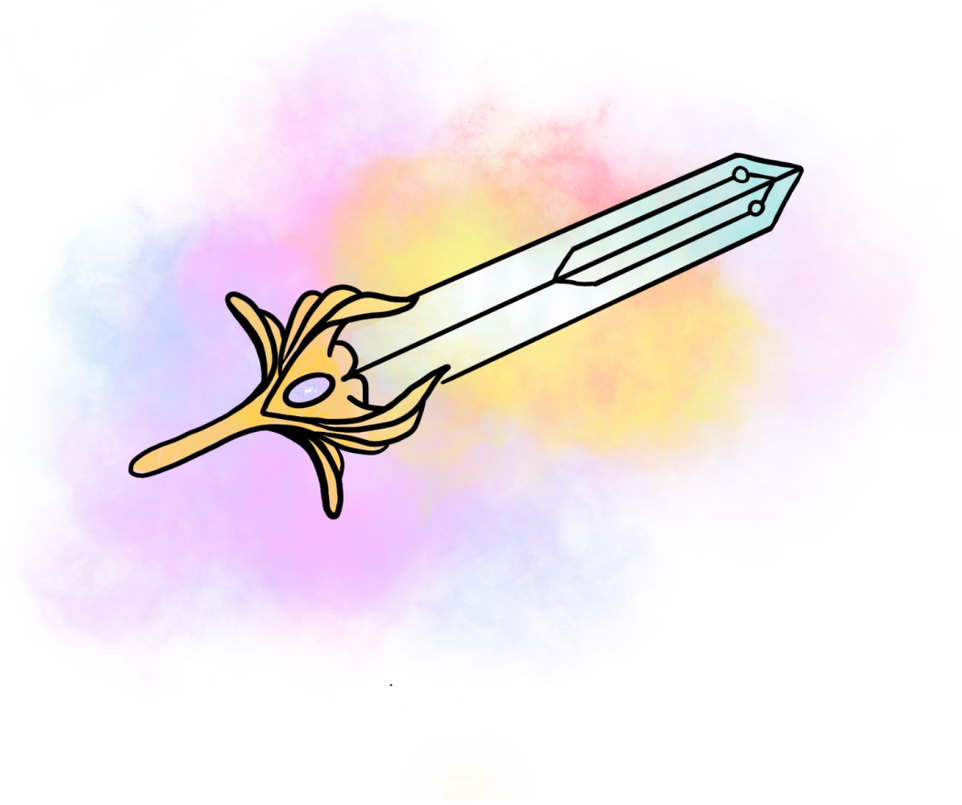
- Ehsan *et al.* promote the concept of *seamful design for XAI:*

*“Instead of hiding seams or treating them as problematic, seamful design argues for strategically revealing (and concealing) seams to support user agency…”*

Ehsan, Upol, et al. "Seamful xai: Operationalizing seamful design in explainable ai." *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1 (2024): 1-29.

# XAI is a double-edged sword

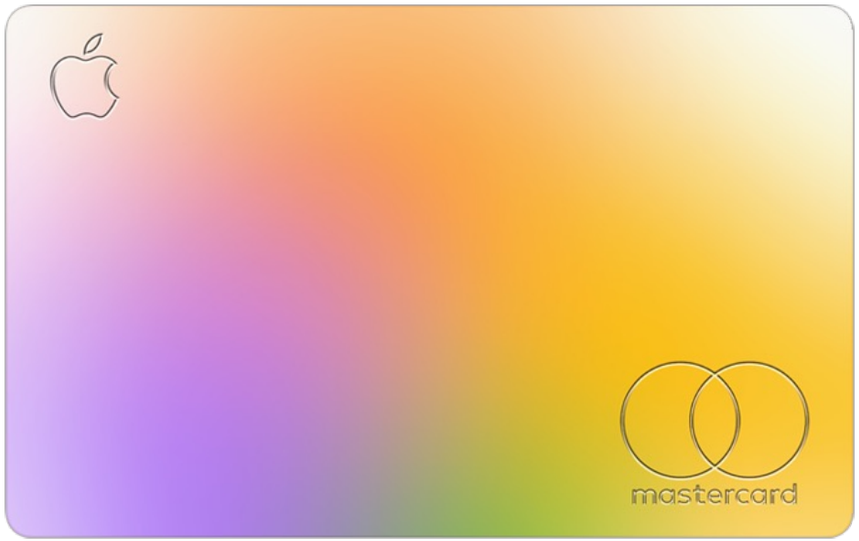
- **“The AI system should be provided in a way that allows the overseer to understand its capabilities and limitations”**
- Different methods of explaining may lead to malicious use of XAI too!
- We are selective over what we choose to explain.
- Explanations can be misleading and misinterpreted, even if all actors have good intentions.



EU AI Act: first regulation on artificial intelligence | Topics | European Parliament (europa.eu).

# Automated Decision-making

- Should offer support for all stakeholders, not just data scientists, but also financial advisors, auditors, policy makers, healthcare providers



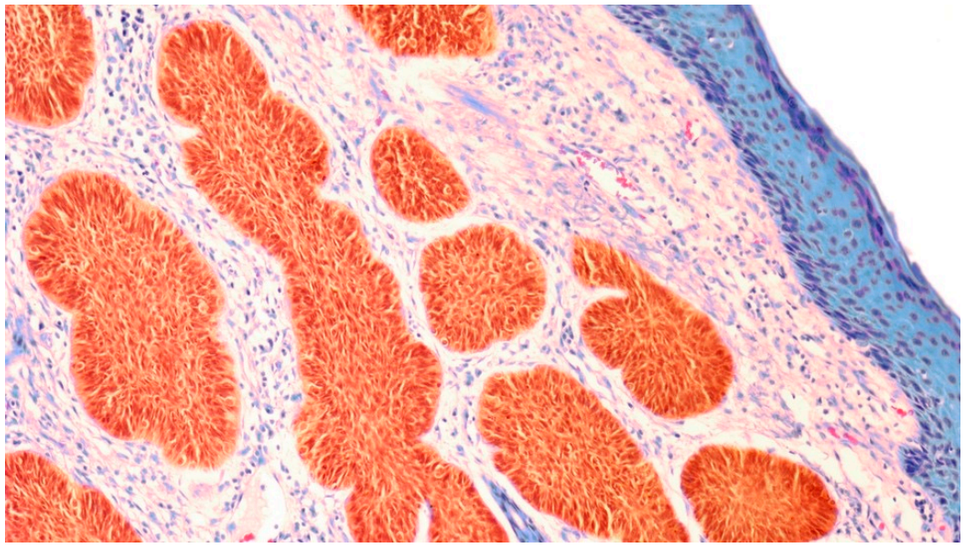



Photo sources: support.apple.com and Steve Gschmeissner / Getty

# Ways forward

- Interdisciplinary methods for impactful XAI methods.
- Human-centricity and context-specificity.
- Interactive and adaptive XAI for effective human-machine teaming.


UMEÅ UNIVERSITY

# Thanks!