

A journey in the land of Explainable AI

(Beautiful landscapes, horrendous pits and everywhere in between)

Julien Girard-Satabin (CEA LIST): julien.girard2@cea.fr, Romain Xu-Darme (CEA LIST)



July 15, 2026

Preparing for the journey

Material

The slides and the companion material are available at https://julien.girard-satabin.fr/teaching/scai_2026/



The lab session can be run in your browser

Myself

1. researcher at CEA on formal methods for software safety and security applied to machine learning;
2. authored several papers on explainable AI and techniques, maintainer of the CaBRNet platform for case-based reasoning;
3. working on case-based reasoning, explainable AI and out-of-distribution detection with academics and industry.



You, the audience

1. graduate students working on AI-adjacent topics;
2. concerned parties curious about the reliability of existing xAI techniques;



I will speak during the course session, but feel free to interrupt

Example definition

When those are on, those are important concept to grasp

Example question

When those are on, it is an open question for you. No wrong answers, just interactions.

Opening questions

1. How are you using AI (generative or not) in your day to day?
2. Where do you think AI (generative or not) is used ?
3. Which guarantees would you expect from an AI system?
4. What is your opinion on accountability for AI systems?

Definitions

Explanation

“An explanation is a presentation of (aspects of) the reasoning, functioning and/or behavior of a machine learning model in human-understandable terms” [Nau+23]

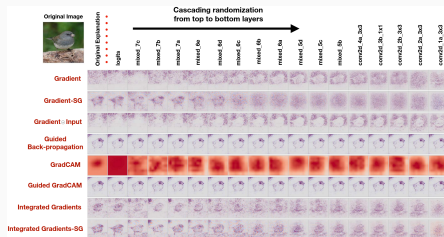
“The **belief** (by the trustor) in the ability (of the trustee) to achieve **something**”

What motivates this course?

Regulations	Technical issues
“Right to explanation” in GDPR [SP17] Lack of thorough and standardized evaluation processes Accountability is key for the adoption of a technology	Article 86 of the AI Act General brittleness of AI models Understanding how society works

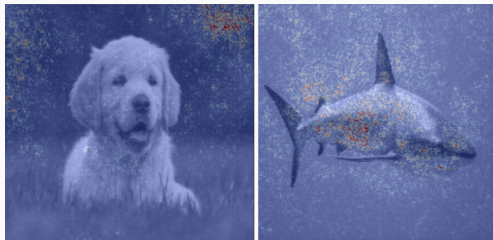
This course: A critical view on the field

Non-fidelity to model behaviour



This course: A critical view on the field

Limited actionability



This course: A critical view on the field

Brittleness



This course: A critical view on the field

Multiple definitions with sometimes conflicting end goals

“Why is this giraffe is classified as an impala” → attribution

“How can I change this image of a shark to be classified as a giraffe” →
contrastive/counterfactual

“This is an apple because it looks like other apples I know” → by correlation

“If this is an accessory fruit from genus Malus Domestica, then this is an apple” →
by causality deduction

This course: A critical view on the field

Explanation evaluation still an open problem

No single-metric dominates: rely on use-case, context dependent evaluations with humans



**Increase stakeholders
trust in the ML system**

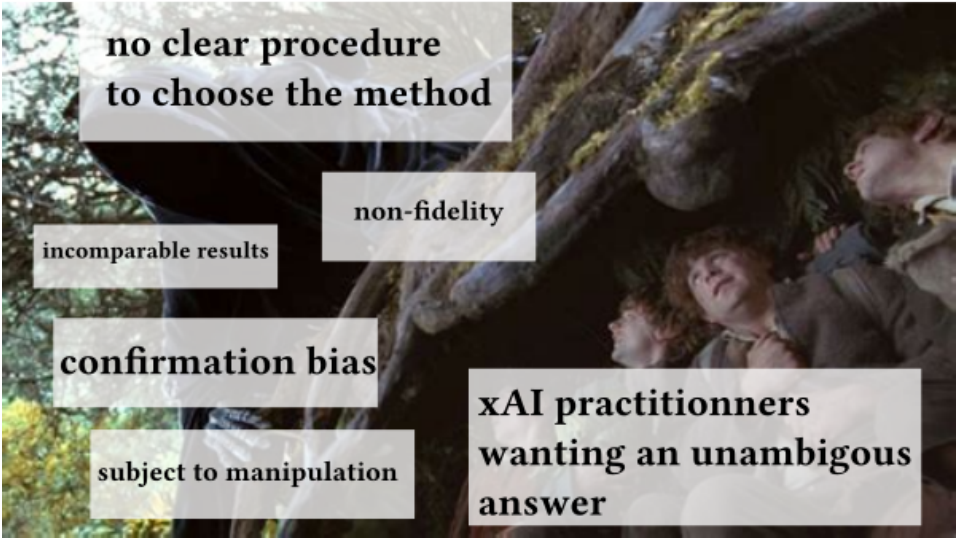
**Tons of paper and
off-the-shelf libraries**

**Debug AI
models at design time**

**Discover the
underlying mathematical
structure of a ML model**

**Reach compliance
with GDPR**

The beautiful landscapes



**no clear procedure
to choose the method**

non-fidelity

incomparable results

confirmation bias

subject to manipulation

**xAI practitionners
wanting an unambiguous
answer**

The monsters to avoid



Everything in between

Topics of interest

Gradient-based attribution methods on neural network image classification, their inner workings, their limitations.

Not covered in this tutorial: symbolic data (see the work of Pierre Marquis), LIME and Shapley values (see [Sla+20; Mar23]), explanations on LLMs and Generative AI (no consensus on the literature on their feasibility/actionability)

On the quality of explanation

Some criterion coming from argumentation theory and psychology on what constitutes a good explanation ([Mil19; AB21]):

1. *contrastive* (why P instead of Q?);
2. *social* (A explains P to B);
3. *generalistic* (cover more facts);
4. *concise* (quote less causes, can also be more *compact*);
5. *coherent* (related to previous knowledge);

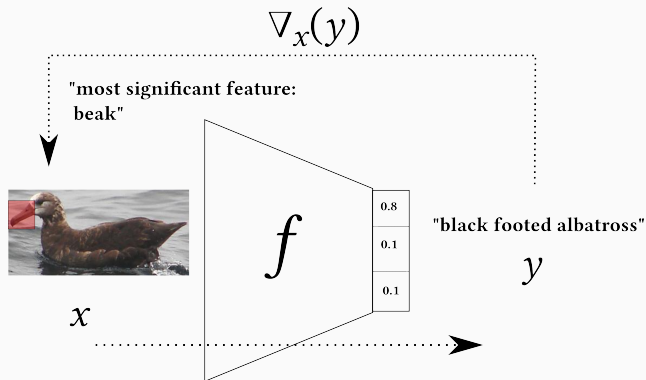
Comparing xAI methods *is still an open research question* [Xu+23a] even more so evaluating their *practical usefulness* [Col+23]

The problems xAI aims to solve

1. developpers: debugging their models at design-time [Gwa+23];
2. external auditors: getting convincing arguments that the program behaves according to a specification or a regulation (GDPR article 13.f [SP17]);
3. end-users: refuting arguments made by a program.

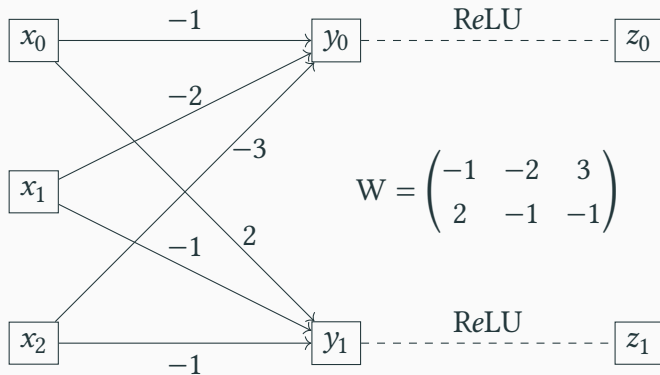
Post-hoc explanations: understanding signs

The general framework of local feature attribution

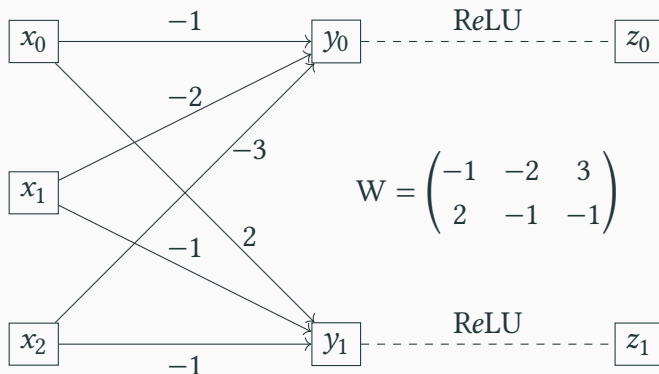


Given an input x , identify the most relevant outputs on $\mathcal{Y}$ and project back on $\mathcal{X}$, obtaining a *heatmap* M

The baseline: gradient computation

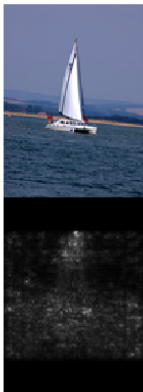


The baseline: gradient computation



For $x_0 = 1, x_1 = -1, x_2 = 0$

$$\begin{aligned} \nabla_x f(x) &= \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x} \\ &= \begin{pmatrix} 1 \\ 3 \end{pmatrix} \cdot W \\ &= \begin{pmatrix} 5 \\ -8 \\ 0 \end{pmatrix} \end{aligned}$$



Saliency Vanilla computation. From [SVZ13]

What are your opinion on those attributions? What is displayed? What is lacking?

Vanilla gradient caveats

G

gradients in high-dimensional spaces tend to have strong, local variations

Removing Noise by Adding Noise

Smoothgrad [Smi+17] $\nabla_{x^*}(y)$ where x^* is a gaussian neighborhood of x

$$\nabla_{x^*}(y) \approx \frac{1}{n} \sum_0^n \nabla_x f(x + \mathcal{N}(0, \sigma))$$

Vannila gradients caveas

G

radient based approaches may not capture subtle variations:

Given $f(x) = 1 - \text{ReLu}(1 - x)$, $\nabla_0 f$ and $\nabla_2 f$ have the same value

Desirable properties

We need guarantees such that if two inputs x and x' results in a different prediction, then explanations of x and x' should differ => **sensitivity**

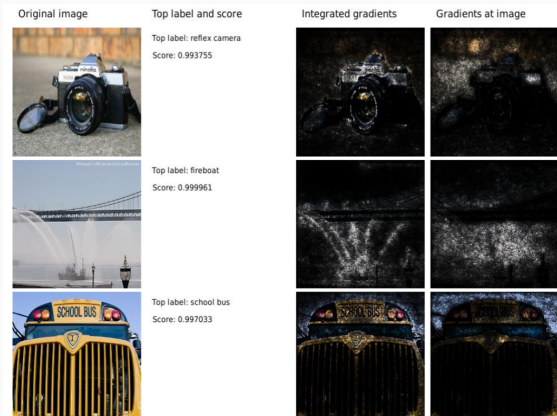
Integrated gradients

Gradient on the line between x and a baseline image x' [STY17]

$$\text{IG}_i = (x_i - x'_i) \int_{\alpha=0}^1 \nabla_{x_i} f(x' + \alpha(x - x')) d\alpha$$

Since it only computes pure gradient between a baseline and an input, the sensitivity property holds

Integrated gradients



From [STY17]

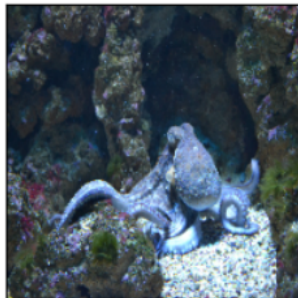
Greyhound (vanilla)



Soup Bowl (vanilla)



Eel (vanilla)



RE 10.9: Images of a dog classified as greyhound, a ramen soup classified as soup bowl, and an octopus classified as eel.

From [Mol22]

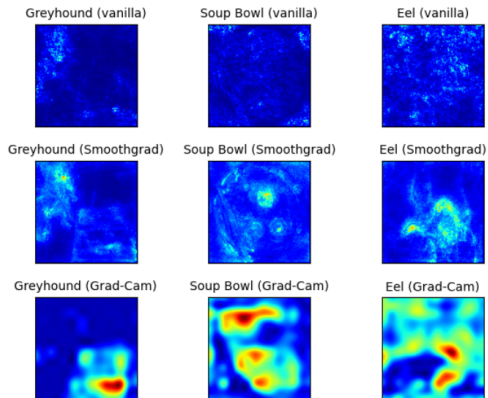


FIGURE 10.10: Pixel attributions or saliency maps for the Vanilla Gradient method, SmoothGrad and Grad-CAM.

From [Mol22]

The network decision is ill-based. *Why* is that? How to fix it?

This explanation does not help to adjust our mental model on the program's behaviour, it is not a good one

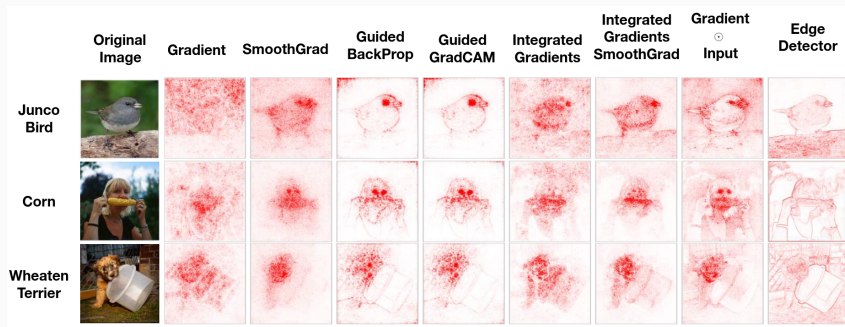
Nuance

Extracting a causal chain and displaying it to a person is causal attribution, not (necessarily) an explanation [Mil19].

Attribution-based approaches are not enough to “fill the holes” for complex programs

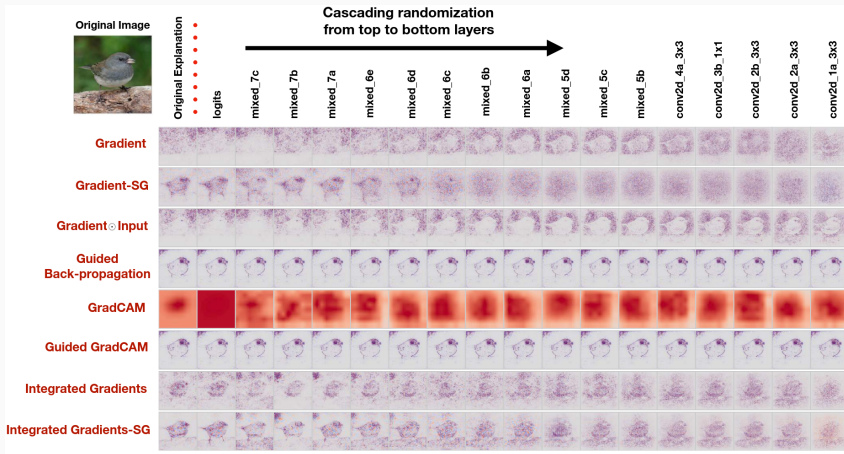
“*How* the decision was taken” and “*Why* the decision was taken” are two different questions

Feeding our own biases



From [Ade+18]

Feeding our own biases



From [Ade+18]. The more on the right, the more random the network is.

Feeding our own biases

Confirmation bias (Wikitionnaire)

(psychology) A cognitive bias towards confirmation of the hypothesis under study

A “nice” heatmap will confirm that the network works as expected, without being necessarily an accurate description of its inner working

Explanations can be manipulated

What if we computed the gradient of the explanation map $\nabla_x M$ to change M such that the provided *explanation* changes?

Explanations can be manipulated



$$\mathcal{L} = \|M(x_{adv}) - M^t\|_p + \lambda \|f(x_{adv}) - f(x)\|_p$$

From [Dom+19]

Explanations can be manipulated... with some limitations

$$\nabla_{x_{adv}} \|M(x_{adv}) - M^t\|_2 \propto \frac{\partial M}{\partial x_{adv}} = \frac{\partial^2 f}{\partial^2 x_{adv}}$$

If f contains ReLU activation functions, we then have

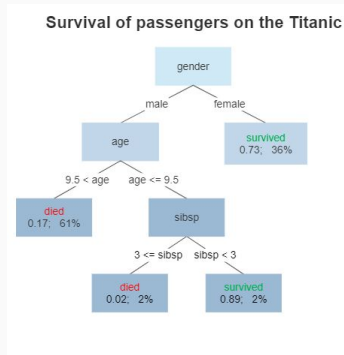
$$\frac{\partial^2 f}{\partial^2 x_{adv}} \propto \text{ReLU}'' = 0$$

Wrapping up: empirical feature attribution approaches

1. usually only require gradient computation access;
2. provide attributions on the input space, but no direct exposition of the underlying decision process;
3. brittle;
4. heavily rely on the program internal representation;
5. no validity domain.

**Explainable by design programs:
mastering the map and the quill**

Decision trees



from Wikipedia https://en.wikipedia.org/wiki/Decision_tree_learning/

Issue: the deeper the tree, the less amenable it is to understand its decision

Integrating decision trees as the decision process for image classification

Case-based reasoning

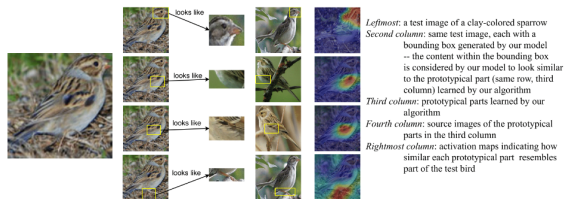
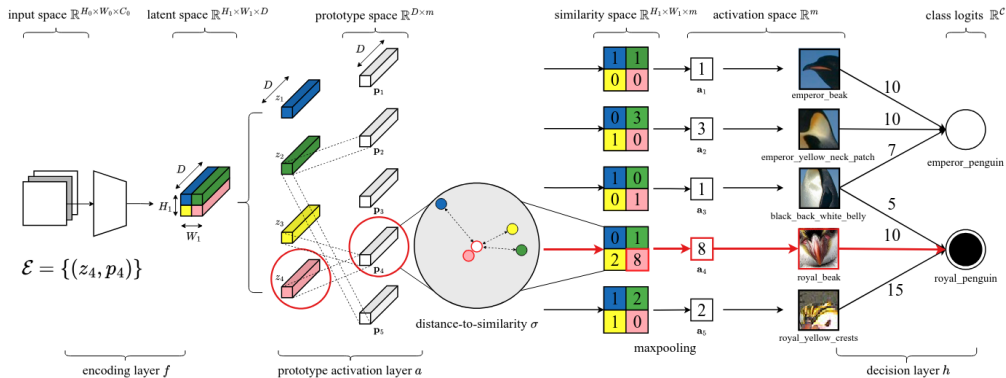


Figure 1: Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species.

From [Che+19]

This looks like that: networks display training set examples related to the new sample



Prototype-based approaches

Concept of *prototype*: an activation vector p_i in the latent space such that it is maximally activated

During inference, they project on that latent space the activation $\mathbf{z}$ for a new input using a projection operator Π

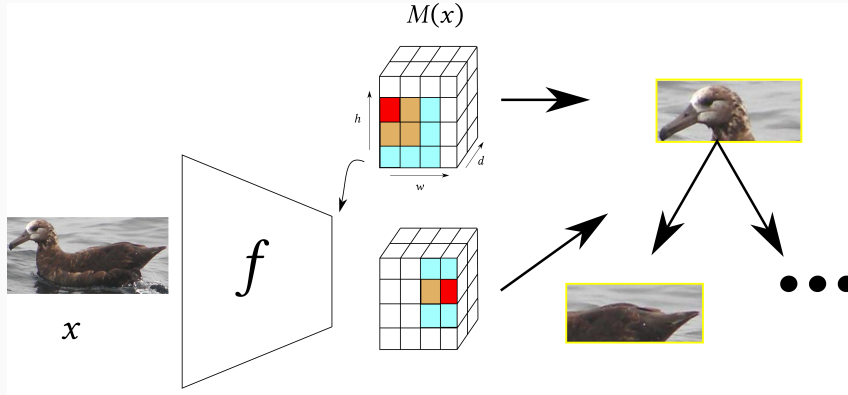
$$\min_i = \Pi(\mathbf{z}, p_i)$$

Afterward, they use a visualisation method v to associate the nearest image x_i in the training set corresponding to p_i .

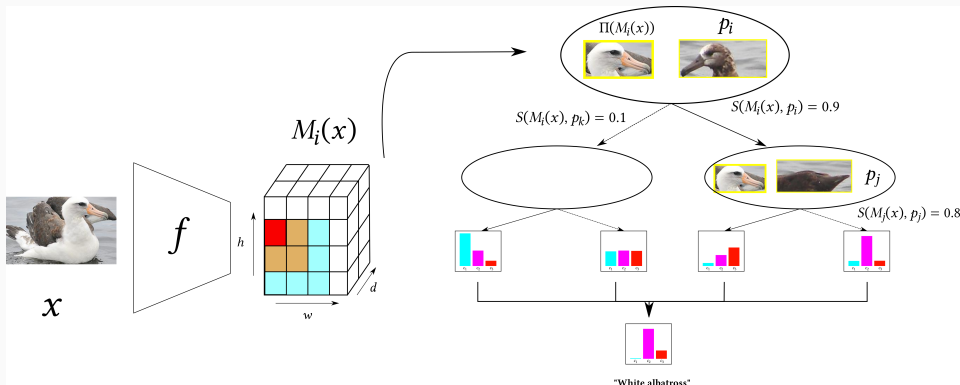
Case-based reasoning ingredients

- Loss function:
 1. a *clustering* loss that constraint the intra-class prototypes to be close to each other (as salient parts in an image are expected to live together);
 2. a *separation* loss that constraint prototypes for a given class to maximize their distance with prototypes from another class.
- Projector Π : the l_2 norm (that is arguably not a good fit in higher-dimensional spaces), the cosine similariy $\frac{x \cdot y}{||x|| ||y||}$
- Visualisation v : choosing an imprecise one can result in a wrong attribution on the input sample and thus, a wrong interpretation.
- p_i : most architectures propose a per-class prototypes number, others allow sharing among classes.

Prototype based approaches - ProtoTrees



Prototype based approaches - ProtoTrees



Prototype based approaches - ProtoTrees

1. learn “prototypes” p : part of the input set that are deemed representative for the prediction;
2. during inference, $M_i(x)$ are compared to p_i using a similarity layer S ;

Protoype based approaches - Tackling tree complexity

ProtoTree have two hyperparameters that influence the decision tree:

1. the decision tree depth;
2. the pruning threshold;

Prototype based approaches - Caveat

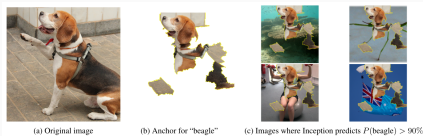
1. Still rely on the hypothesis that similarity in the feature space equals similarity in the input space;
2. Need retraining models;

Next trails to uncover for xAI

A validity domain for feature-attribution methods

Given f , x and an explanation $M_{f(x)}$, what happens for x' ? For f' ?

Potential inspirations: Anchors [RSG18], Formal Abductive Explanations based on perturbations [Wu+24]



Evaluating the quality of explanations

How to evaluate explanations and sort evaluation metrics? See [Xu-+23a; Xu-+23b]: there is no “one size fits all” metric.

Is relying on a *metric* even sensible when we explain to humans? Bridges to be done with psychology and Human-Machine Interface [AAA24]

Correctness regarding the model

How using formal methods to ensure the correctness of the method [MI22]?

See our work [Sor+25; Don+25]!

Social explanation

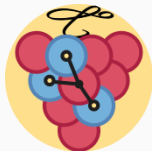
“Social” explanation is yet to happen [Oue22]

User-centric evaluations must be developed [Col+23]

The field must come up with its own paradoxes [Afr+26]

Join us for the rest of the journey!

1. explainable by design design and evaluation with CaBRNet [Xu+24];
2. formal verification of AI [Alb+22];



<https://www.caisar-platform.com>

Legal notice

Introduction images are drawn from Lord of the Ring (Peter Jackson), Elden Ring (From Software)

References

- [AAA24] Noor Al-Ansari, Dena Al-Thani, and Reem S. Al-Mansoori. **“User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review”**. In: *Human Behavior and Emerging Technologies* 2024.1 (2024), p. 4628855. ISSN: 2578-1863. DOI: 10.1155/2024/4628855. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2024/4628855> (visited on 01/29/2026) (cit. on p. 59).

Bibliography ii

- [AB21] Katie Atkinson and Trevor Bench-Capon. “**Argumentation Schemes in AI and Law**”. In: *Argument & Computation* 12.3 (2021), pp. 417–434. ISSN: 1946-2174. DOI: 10.3233/AAC-200543 (cit. on p. 20).
- [Ade+18] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. “**Sanity Checks for Saliency Maps**”. In: *Advances in Neural Information Processing Systems* 32. 2018, p. 11 (cit. on pp. 38, 39).

Bibliography iii

- [Afr+26] Saleh Afroogh, Seyd Ishtiaque Ahmed, Petra Ahrweiler, David Alvarez-Melis, Mansur Maturidi Arief, Emilia Barakova, Falco J. Bargagli-Stoffi, Erdem Biyik, Hanjie Chen, Xiang 'Anthony' Chen, Robert Alan Clements, Keeley Crockett, Amit Dhurandhar, Fethiye Irmak Dogan, Mollie Dollinger, Motahhare Eslami, Aldo A. Faisal, Arya Farahi, Melanie F. Pradier, Saadia Gabriel, Diego Garcia-Olano, Marzyeh Ghassemi, Shaona Ghosh, Hatice Gunes, Ehsan Hajiramezanali, Stefan Haufe, Biwei Huang, Angel Hwang, Md Tauhidul Islam, Junfeng Jiao, Amir-Hossein Karimi, Saber Kazeminasab, Anastasia Kuzminykh, William La Cava, Brian Y. Lim, Xiaofeng Liu, Mohammad R. K. Mofrad, Alicia Parrish, Maria Perez-Ortiz, Shriti Raj, Swabha Swayamdipta, Salmon Talebi, Kush R. Varshney,