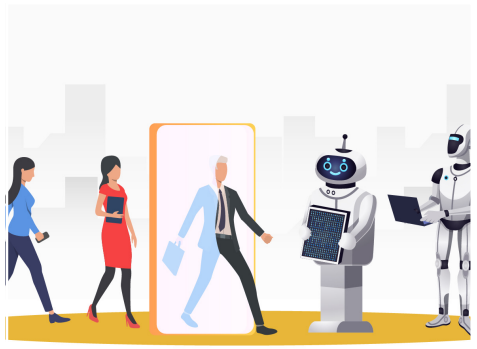
Hiccups on the Road to XAI

Security and Privacy risks of Explainable AI

Ulrich Aïvodji
Assistant Professor - ETS
ulrich.aivodji@etsmtl.ca

ML in high-stakes decision

Healthcare



Hiring

Finance

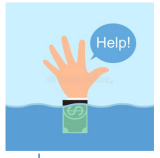
Need transparency

Why do we need transparency ?

Compliance

Debugging

Fairness



Recourse

Regulatory agencies

Engineers & researchers

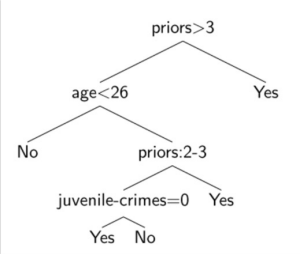
End-users

Achieving transparency

Transparency by design

1.	*Prior Arrests ≥ 2*	1 point		···
2.	*Prior Arrests ≥ 5*	1 point	+	···
3.	*Prior Arrests for Local Ordinance*	1 point	+	···
4.	*Age at Release between 18 to 24*	1 point	+	···
5.	*Age at Release ≥ 40*	-1 point	+	···
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	···

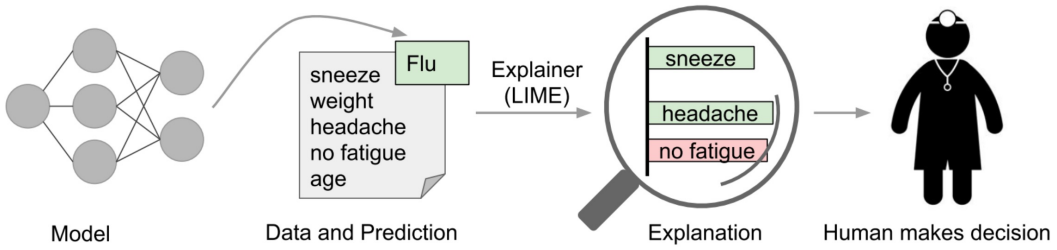
SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%



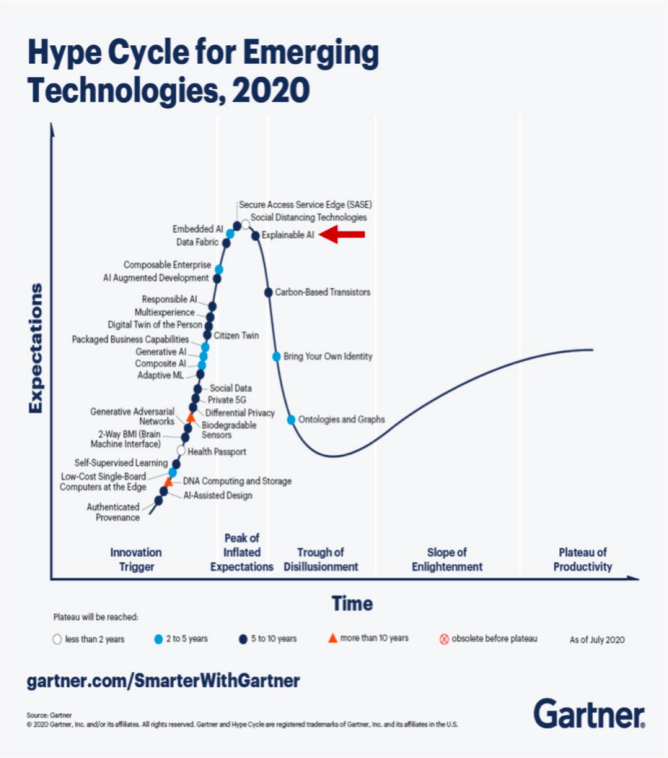
$$f(\mathbf{x};\boldsymbol{\theta}) = b + w_1 x_1 + \cdots + w_D x_D = b + \mathbf{w}^\top \mathbf{x}$$

if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

Post-hoc explanation



Explainable AI



Explainable AI

A Survey of Methods for Explaining Black Box Models

RICCARDO GUIDOTTI, ANNA MONREALE, SALVATORE RUGGIERI, and
FRANCO TURINI, KDDLab, University of Pisa, Italy
FOSCA GIANNOTTI, KDDLab, ISTI-CNR, Italy
DINO PEDRESCHI, KDDLab, University of Pisa, Italy

In recent years, many accurate decision support systems have been constructed as black boxes, that is as systems that hide their internal logic to the user. This lack of explanation constitutes both a practical and an ethical issue. The literature reports many approaches aimed at overcoming this crucial weakness, sometimes at the cost of sacrificing accuracy for interpretability. The applications in which black box decision systems can be used are various, and each approach is typically developed to provide a solution for a specific problem and, as a consequence, it explicitly or implicitly delineates its own definition of interpretability and explanation. The aim of this article is to provide a classification of the main problems addressed in the literature with respect to the notion of explanation and the type of black box system. Given a problem definition, a black box type, and a desired explanation, this survey should help the researcher to find the proposals more useful for his own work. The proposed classification of approaches to open black box models should also be useful for putting the many research open questions in perspective.

(ACM Computing Surveys. 2018)

- Model (global) explanation techniques
- Outcome (local) explanation techniques

Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI

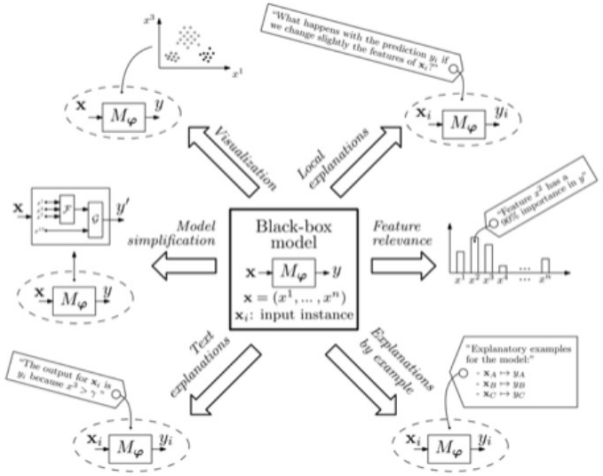
Alejandro Barredo Arrieta[a], Natalia Díaz-Rodríguez[b], Javier Del Ser[a,c,d], Adrien Bennetot[b,e,f], Siham Tabik[g], Alberto Barbado[h], Salvador Garcia[g], Sergio Gil-Lopez[a], Daniel Molina[g], Richard Benjamins[h], Raja Chatila[f], and Francisco Herrera[g]

[a]TECNALIA, 48160 Derio, Spain
[b]ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
[c]University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain
[d]Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain
[e]Segula Technologies, Parc d'activité de Pissaloup, Trappes, France
[f]Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, France
[g]DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain
[h]Telefonica, 28050 Madrid, Spain

Abstract

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of the latest techniques brought by sub-symbolism (e.g. ensembles or Deep Neural Networks) that were not present in the last hype of AI (namely, expert systems and rule based models). Paradigms underlying this problem fall within the so-called *eXplainable* AI (XAI) field, which is widely acknowledged as a crucial feature for the practical deployment of AI models. The overview presented in this article examines the existing literature and contributions already done in the field of XAI, including a

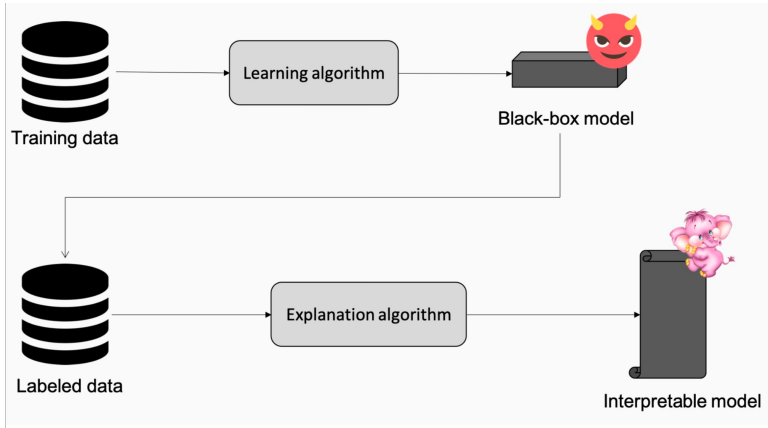
(arXiv:1910.10045v1. 2019)



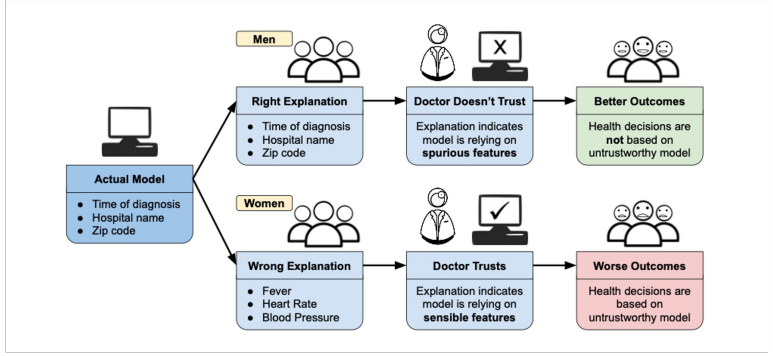
XAI Challenges

Trust

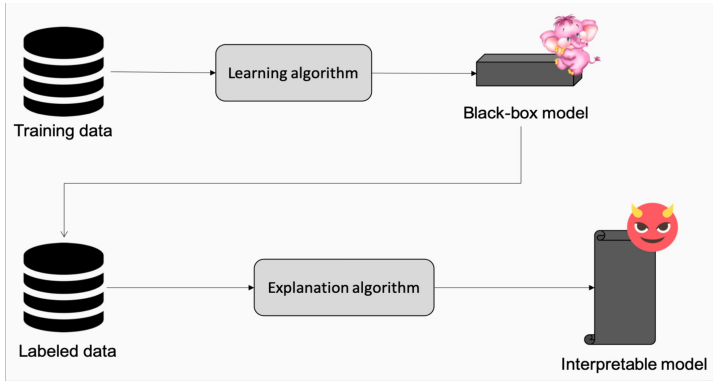
- Auditor
- End-User

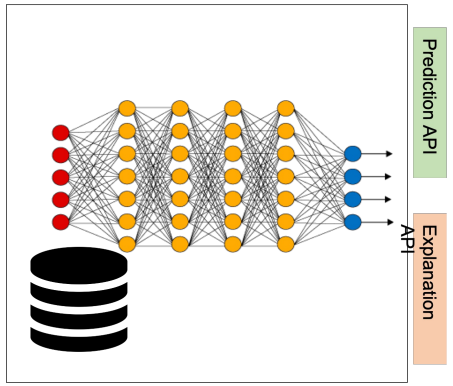


- Practioner



Security





Service provider

Trust challenges

Manipulation risks

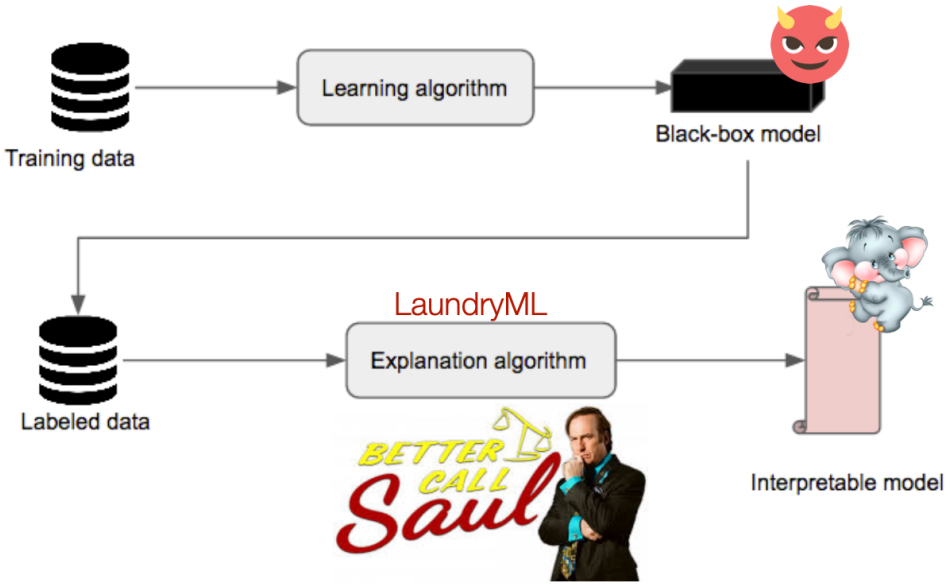
Fooling auditors and end-users

Fairwashing

Dishonest ML models' producers could leverage on post-hoc explanation techniques to perform fairwashing

↓

Given the false perception that a ML model complies with a given ethical value



We consider fairness as the ethical value to "fairwash"

Fairwashing: the risk of rationalization

Ulrich Aïvodji[1], Hiromi Arai[2,3], Olivier Fortineau[4], Sébastien Gambs[1], Satoshi Hara[5], Alain Tapp[6,7]

1) Universite du Québec à Montréal
2) RIKEN AIP
3) JST PRESTO
4) ENSTA ParisTech
5) Osaka University
6) Université de Montréal
7) MILA

Social demands on ethically aligned AI

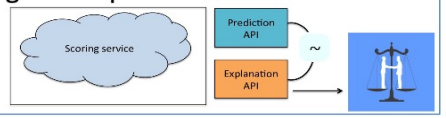
Respect Human Rights
- "Right to an explanation"
- "Right to fairness and transparency"

Explanation and Fairness are required for ML.

"Fairwashing"

Faking as if ML model is ethically aligned.

This Study
- Raise the awareness on the risk of fairwashing.
- Show that an unfair model can be rationalized by generating fake explanations.



LaundryML

Systematically generating fake explanations → Unfair decisions can be rationalized.

LaundryML

[Step1] Generate many explanation by model enumeration.

[Step2] Pick up one that is useful for "fairwashing".

- Enumeration of CORELS

Enumerate models w/ small misclassification and unfairness.

$$\text{obj}(\cdot) = (1-\beta)\text{misc}(\cdot) + \beta\text{unfairness}(\cdot) + \lambda K$$

Two building blocks:

1. CORELS [Angelino+, KDD'17]
   - Exact learning algorithm for rule list.
2. Model Enumeration [Hara+,AAAI'17,'18]
   - Enumerate models in ascending order of the training errors.

Fidelity / Misclassification

Similarity/Dissimilarity between the base model *b* and the explanation *c*.

$$\text{fidelity}(c) = \frac{1}{|X|}\sum_{x\in X}\mathbb{I}(c(x)=b(x)),\ \text{misc}(c) = 1-\text{fidelity}(c)$$

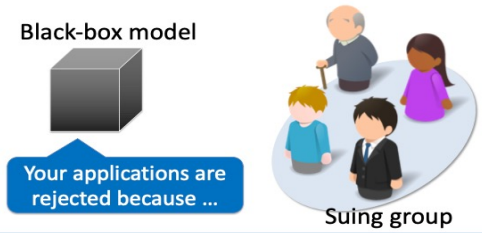
Unfairness (demographic parity)

A bias in the model's decision.

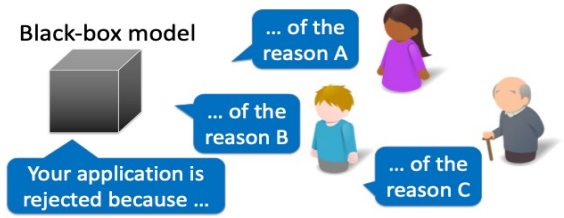
$$\text{unfairness}(c) = |P(\hat{y}=1|s=1) - P(\hat{y}=1|s=0)|$$

Variants of LaundryML

LaundryML-global (Model rationalization)
- Rationalize decisions on the group (the suing group)
- Generate a single fake explanation.



LaundryML-locoal (Outcome rationalization)
- Rationalize individual decisions.
- Generate fake explanations for each individual.



Experiments

Datasets

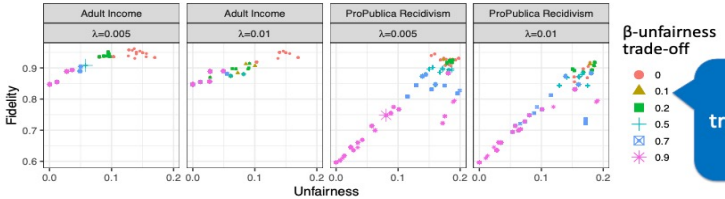
Adult Income (*s*: gender) / ProPublica Recidivism dataset (*s* : race)

→ three subsets: the training set, the suing group, the test set

Classifier Random Forest

Model rationalization by LaundryML-global

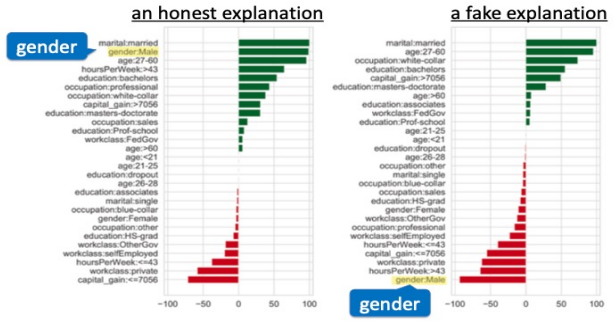
Fidelity and unfairness of rationalized explanations on the suing group



Fidelity – Unfairness tradeoffs in enumerated explanations

Evaluation by FairML

Obtained relative feature dependence ranking (Adult)
- An obtained fake explanation successfully rationalized the unfair decisions.



Fake Explanation

If capital gain > 7056 **then** high-income

else if marital = single **then** low-income

else if education = HS-grad **then** low-income

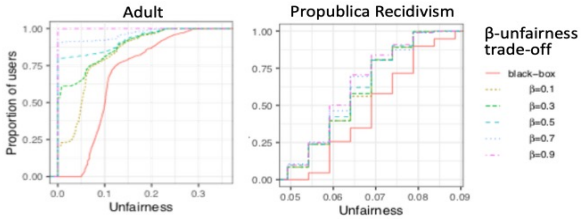
else if occupation = other **then** low-income

else if occupation = white-colloar **then** high-income

else low-income

Outcome rationalization by LaundryML-local

Empirical CDF of the unfairness of the best model



Individual decisions were successfully rationalized.

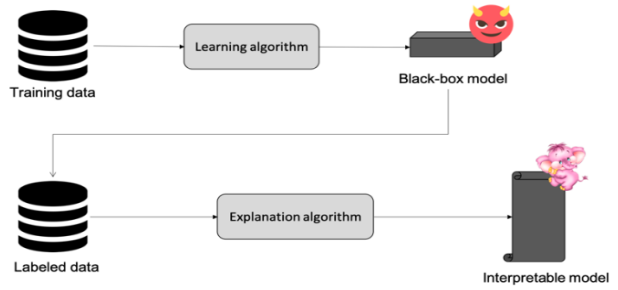
For ProPublica, explanations for each individual were at least twice less unfair than the black-box model.

Characterizing the risk of fairwashing

Ulrich Aïvodji
Hiromi Arai
Sébastien Gambs
Satoshi Hara

Introduction

- Post-hoc explanations are vulnerable to manipulations
- Can we detect these manipulation ?
- We focus on **fairwashing**.



– Given a black-box model b, under-report the unfairness of b, using high-fidelity explainers e.

$$\text{minimize } \mathbb{E}_{(x_i,b(x_i))\sim\mathcal{D}_{sg}}[l(e(x_i), b(x_i))], \quad \text{subject to } \mathtt{unf}_{\mathcal{D}_{sg}}(e) \leq \epsilon$$

* $l(e(x), b(x))$: loss function
* $\mathcal{D}_{sg} = \{X_{sg}, b(X_{sg})\}$: suing group
* $\mathtt{unf}_{\mathcal{D}_{sg}}(\cdot)$: unfairness
* ϵ: value of the unfairness constraint.

– Fidelity is defined as label agreement between b and e

$$\text{fidelity}(e) = \frac{1}{|X|}\sum_{x\in X}\mathbb{I}(e(x) = b(x)).$$

Methods

- Evaluate the **detectability** of fairwashing attacks through the inconsistencies introduced by dataset update or model update on their performances (*i.e.*, fidelity of the fairwashed explainer).
  – **Generalization** of fairwashing attacks beyond suing groups.
  – **Transferability** of fairwashing attacks across black-box models.

- **Fairwashing attacks** have high **manipulability** (*i.e.*, given an unfair black-box model, it is easy to find high-fidelity explainers with significantly lower unfairness). They can **generalize** beyond explanation instances and **transfer** across black-box models.

- *Fairwashing attacks* are significantly **less expensive** to perform than previously thought. **Detecting** fairwashing based on the change of fidelity alone is not a **viable solution** (or at least it is very difficult)

- There is a hope: computing the **range of the unfairness** of high-fidelity explainers can help in **quantifying** the risk of fairwashing.

Take a picture to download the full paper

- **Quantify** the risk of fairwashing by evaluating the range of the unfairness of high-fidelity explainers, using the `Fairness In The Rashomon Set` (Coston *et al.*, 2021) framework:

$$\text{minimize } \mathtt{unf}_{\mathcal{D}_{sg}}(e), \quad \text{subject to } \mathbb{E}_{(x_i,b(x_i))\sim\mathcal{D}_{sg}}[l(e(x_i), b(x_i))] \leq v$$

Experiments

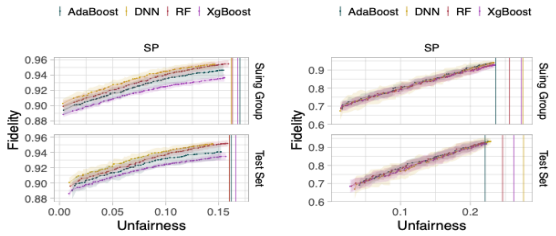


Figure 1: Generalization of fairwashing attacks. Adult Income (Left), COMPAS (Right).

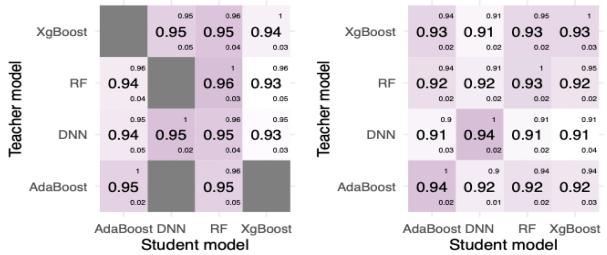


Figure 2: Transferability of fairwashing attacks. Adult Income (Left), COMPAS (Right).

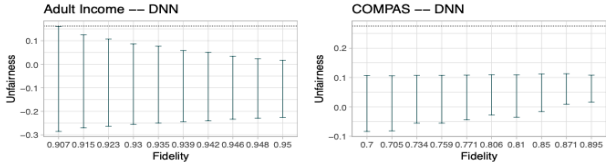


Figure 3: Quantifying fairwashing. Adult Income (Left), COMPAS (Right).

Key takeaways

- Post-hoc explanations techniques can be used to under-report the discrimination level of black-box models that they explain
- Both global and local explanations techniques can be used to achieve this goal.
- It seems almost impossible to detect when fairwashing is performed

Don't let yourself be fooled by post-hoc explanation techniques

More info

Fairwashing: the risk of rationalization
Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, Alain Tapp. (ICML, 2019)

Characterizing the risk of fairwashing
Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, Satoshi Hara. (NeurIPS, 2021)

Washing The Unwashable: On The (Im)possibility of Fairwashing Detection
Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aïvodji, Aisha Alaagib Alryeh Mkean, Sébastien Gambs, Nicolas Papernot. (NeurIPS, 2022)

Adversarial XAI: https://github.com/hbaniecki/adversarial-explainable-ai
by Hubert Baniecki

Countermeasures

- Quantify the risk by analyzing the Rashomon set (set of good models)
  - Range of unfairness (or other useful properties)

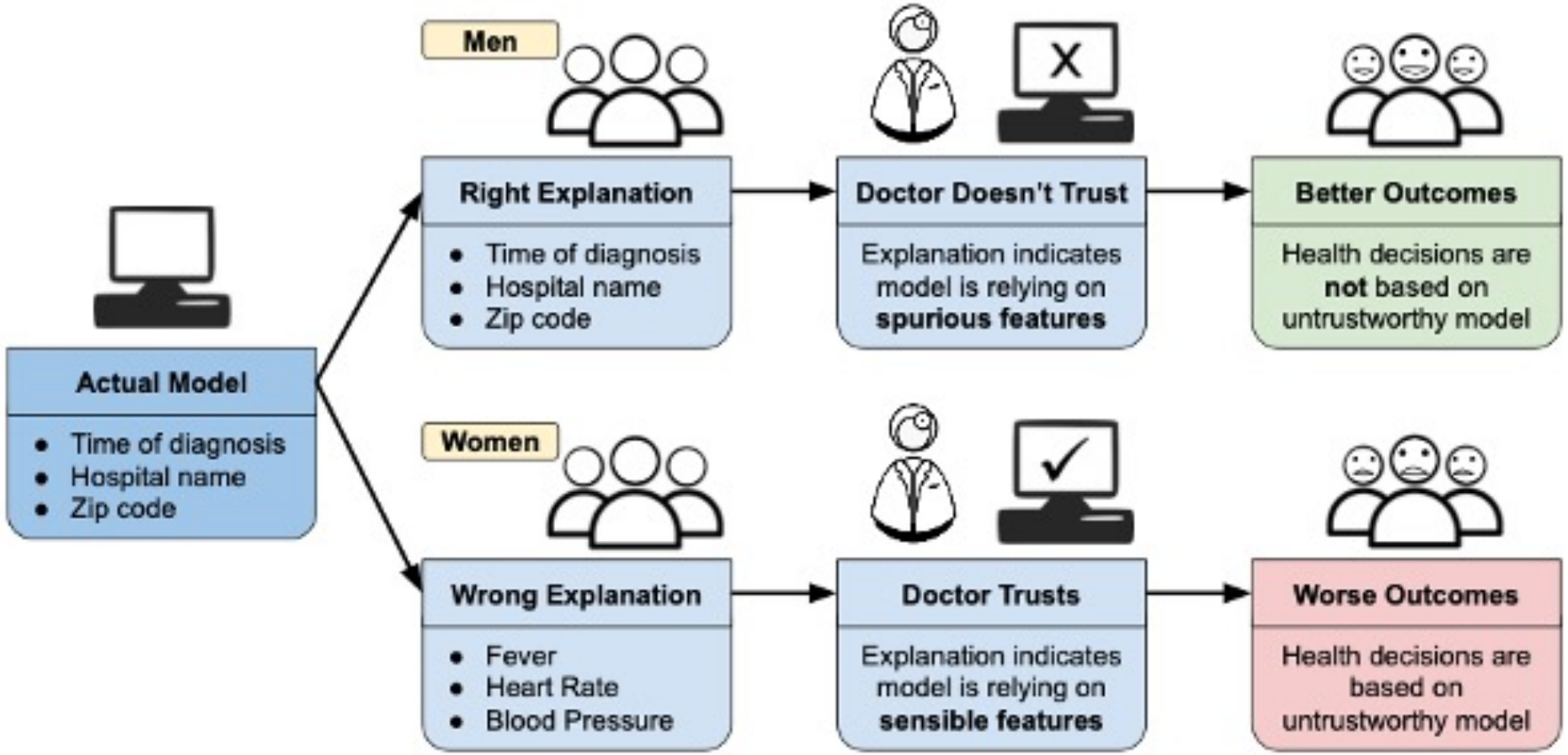
- **Pros**
  - Help in assessing the manipulability of explanations

- **Cons**
  - **Computational overhead**
  - **Some promising directions**
    - Computing the Collection of Good Models for Rule Lists (Mata, Kanamori, Arimura, 2022)
    - Predictive Multiplicity in Classification (Marx, Calmon, Ustun, 2021)
    - Characterizing Fairness Over the Set of Good Models Under Selective Labels (Coston, Rambachan, Chouldechova, 2021)
    - Partial order: Finding Consensus among Uncertain Feature Attributions (Laberge, Pequignot, Khomh, Marchand, Mathieu, 2021)

Subgroup performance disparities

Fooling practioners

Implication of subgroup performance disparities



In this work

Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations.
Jessica Dai, Sohini Upadhyay, Ulrich Aïvodji, Stephen Bach, Hima Lakkaraju. (AIES, 2022)

Evaluation framework to quantitatively measure group-based disparities in the quality of explanations output by state-of-the-art local explanation methods

Overview

Explanation techniques

- LIME
- SHAP
- MAPLE
- SmoothGrad
- IntGrad
- VanGrad

Performance metrics

- Ground truth fidelity
- Prediction gap
- Stability
- Consistency
- Sparsity

Results

		LIME	SHAP	SmoothGrad	IntGrad	VanillaGrad	Maple
German Credit	LR	2	2	1	2	2	0
	NN	0	0	1	1	1	0
Student Performance	LR	0	0	0	1	0	0
	NN	0	2	0	2	0	1
COMPAS	LR	1	2	0	1	1	0
	NN	0	2	1	2	2	0

Table 2: Here we aggregate the instances of significant explanation disparity reported in Table 1 across metrics, counting the number of times significant explanation disparity occurs in the explanation/model/dataset combinations. Across these 36 combinations, explanation disparity occurs in at least one metric 56% of the time.

Key takeaways

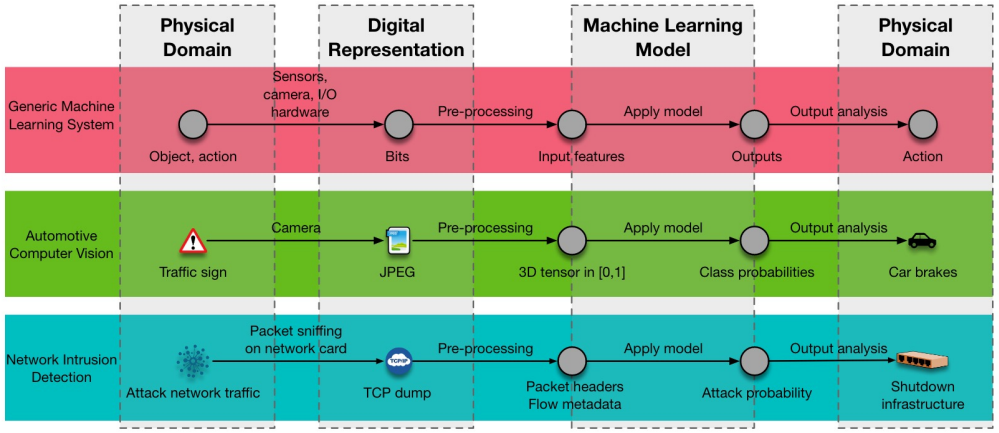
- Post-hoc explanations exhibit **subgroup performance disparities**.

- **What can we do about it?**
  - Study when disparities in explanation quality arise
  - Develop explanation techniques not susceptible to such disparities

Security challenges

Security model

- Attack Surface



- Actors

Data owners

Model developers

End-users

External

SoK: Security and privacy in machine learning. (Papernot *et* al., 2018)

Attack objectives

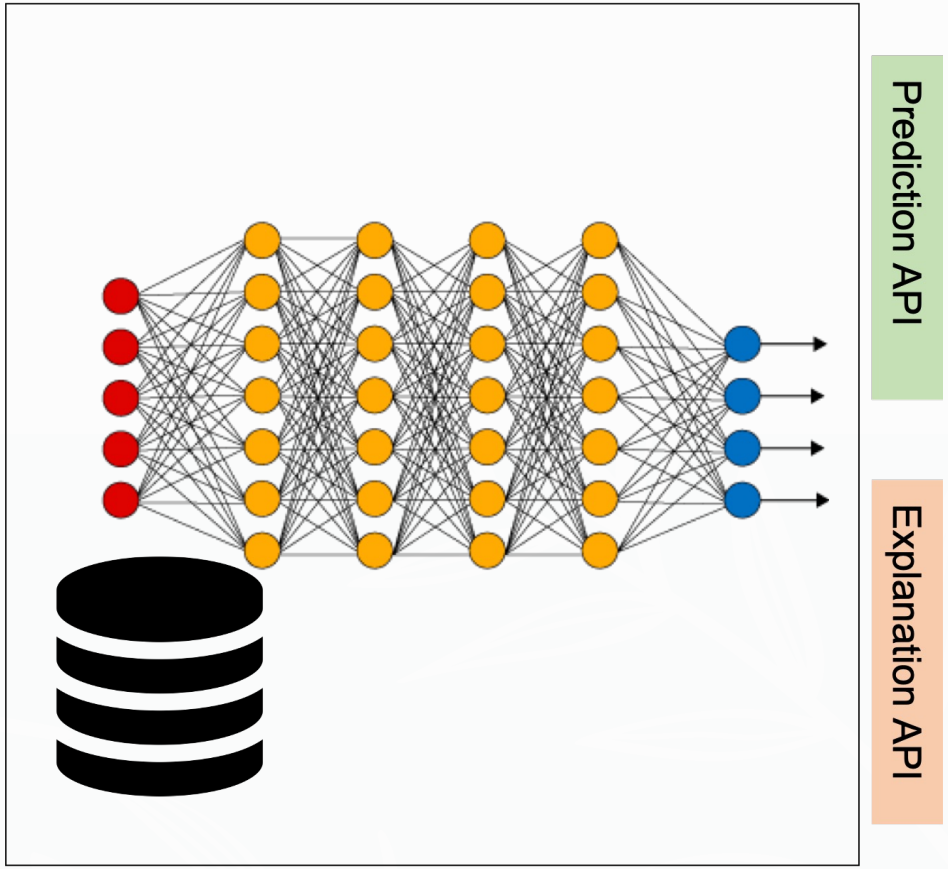
Confidentiality	• Extracting information about the data or the model
Integrity	• Control (targeted or not) of the model prediction
Availability	• Deny of service

Type of Attacks

	Confidentiality	**Integrity**	**Availaibility**
Training		Data poisoning Model poisoning	Sponge poisoning
Inference	Membership inference Property inference Model inversion Model reconstruction	Adversarial examples	Sponge examples

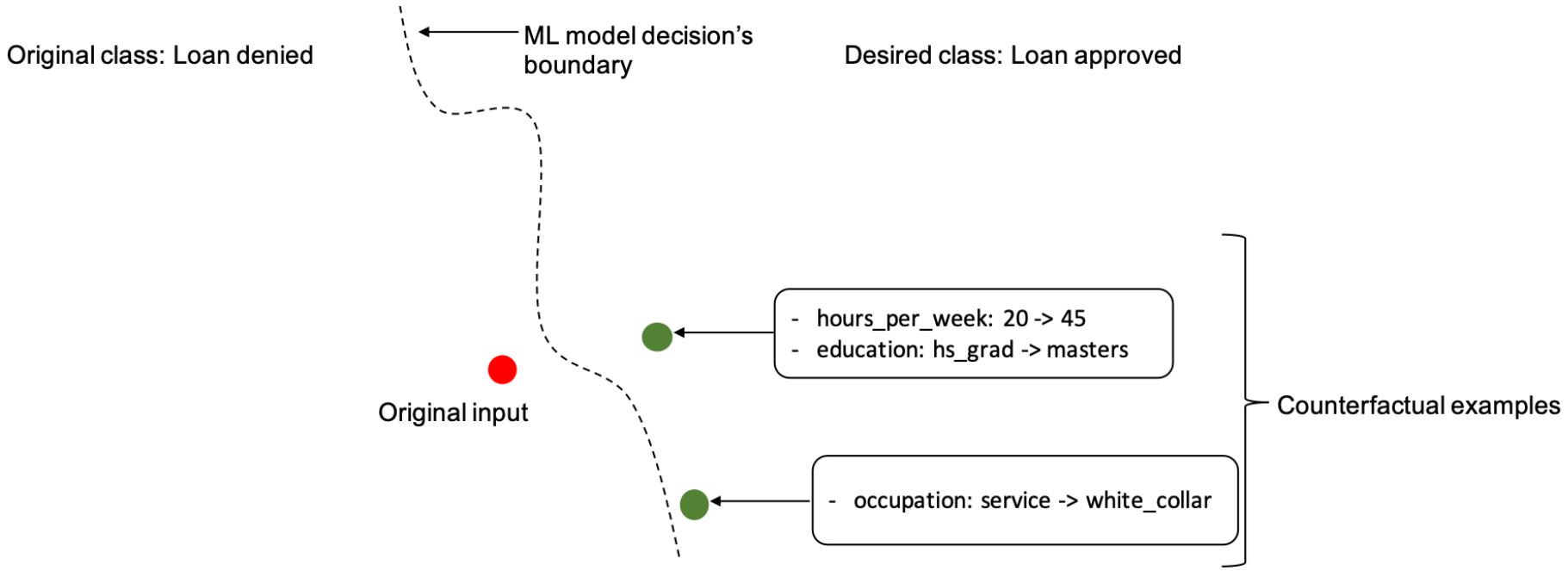
Explanation-based attacks

Case study: counterfactual explanation



Model extraction from counterfactual explanations. (Ulrich Aïvodji, Alexandre Bolot, Sébastien Gambs, 2020)

Counterfactual explanation



Counterfactual explanation

Definition (Wachter et *al.*, 2017)

Given a black-box model b, an original input x_0, its predicted outcome $y_0 = b(x_0)$ and a desired outcome $y \neq y_0$, a counterfactual explanation $c(x_0)$ for the input x_0 is the solution of: $c(x_0) = \underset{c}{\mathtt{argmin}}\, yloss(b(c), y) + |c - x_0|$

DiCe (Mothilal et *al.*, 2020)

$$C(x_0) = \underset{c_1,\ldots,c_k}{\mathtt{argmin}} \frac{1}{k}\Sigma_{i=1}^{k} yloss(b(c_i), y) + \frac{\lambda_1}{k}\Sigma_{i=1}^{k}|c_i - x_0| - \lambda_2 \mathtt{dpp_diversity}(c_1, \ldots, c_k),$$

Counterfactual-based model extraction

Definition

Given a target model b, its prediction API $b(\cdot)$ and its counterfactual explanation API $C(\cdot)$, both available in a black-box setting, a set of data points $x_1, \ldots, x_n$, build a surrogate $S_A \approx b$, using an attack process $\psi(\cdot)$, $b(x_i)$, and $C(x_i)$, with $i = 1, \ldots, n$.

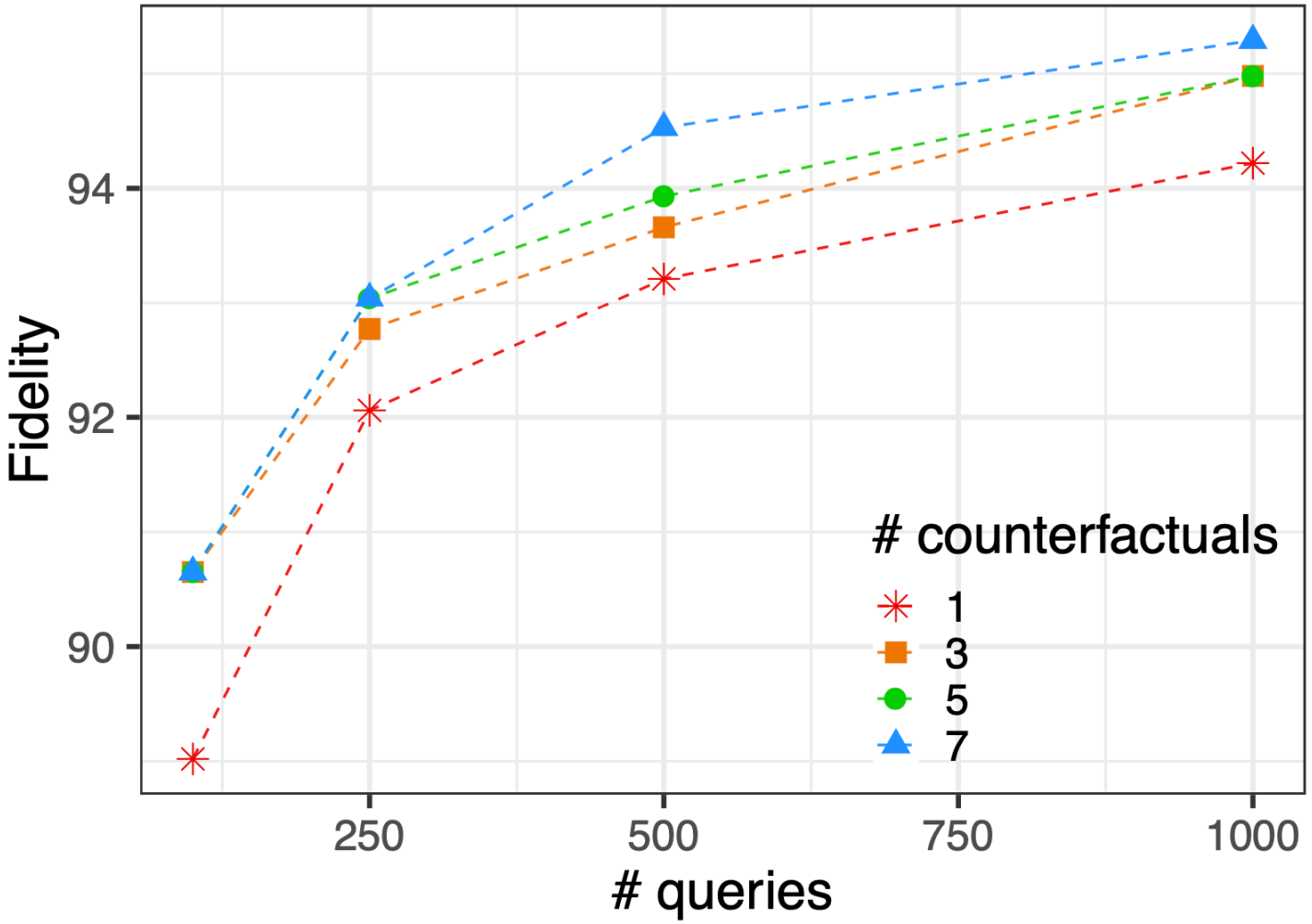
Experiments

- 3 datasets: Adult Income, COMPAS, Default Credit
- Data distribution: known (S1), marginal distributions (S2), unknown (S3)
- Target architecture: known, unknown
- Training data support: used, not used
- Number of queries: 100, 250, 500, 1000
- # data instances in attack pools:
  - Adult Income: 8059
  - COMPAS: 1192
  - Default Credit: 4948

CF-based model extraction for known data distributions

Dataset	Target model Architecture	Training data used by $C(\cdot)$	100 Queries	250 Queries	500 Queries	1000 Queries	Baseline Model
Adult Income	known	yes	89.02/81.05	92.06/82.94	93.21/83.26	94.22/83.68	81.28/76.06
		no	89.39/81.47	91.78/82.87	92.17/82.74	94.84/83.88	
	unknown	yes	89.27/81.11	92.42/83.18	93.62/83.52	94.65/83.88	81.28/76.06
		no	88.99/81.24	92.21/83.05	93.40/83.28	94.89/83.97	
COMPAS	known	yes	87.13/66.19	91.29/67.30	92.17/67.11	92.85/66.97	71.42/61.09
		no	87.91/65.81	89.57/65.86	92.62/66.49	93.92/66.50	
	unknown	yes	88.13/66.50	90.81/67.26	92.00/67.16	92.36/66.90	85.04/64.95
		no	89.12/66.16	90.08/66.03	92.91/66.66	93.43/66.49	
Default Credit	known	yes	97.09/80.22	97.93/80.55	98.31/80.63	98.57/80.52	88.52/77.86
		no	97.15/80.20	97.77/80.34	97.77/80.34	98.28/80.48	
	unknown	yes	97.08/80.12	97.99/80.57	98.39/80.58	98.39/80.58	88.52/77.86
		no	96.90/80.15	97.52/80.38	97.90/80.4	98.03/80.43	

CF-based model extraction for multiple counterfactuals



Key takeaways

- Counterfactual explanations can be exploited to conduct high-fidelity and high-accuracy model extraction attacks under low query budgets
- Making explanations more realistic favors stronger model extraction attacks

Privacy-preserving XAI

- Model Explanations with Differential Privacy. (Neel Patel, Reza Shokri, Yair Zick, 2022)
- Potentially helps prevent membership inference
- Not model reconstruction attacks

Thank you for your attention!