Algorithmic Recourse and Explainable Counterfactual Interventions

Giovanni De Toni

giovanni.detoni@unitn.it
@giovanni_detoni

Mobile and Social Computing Lab, FBK, Italy
Structured Machine Learning Group, University of Trento, Italy

25th February 2022

User

"Unfortunately, we cannot offer you any loan"

Decision Maker

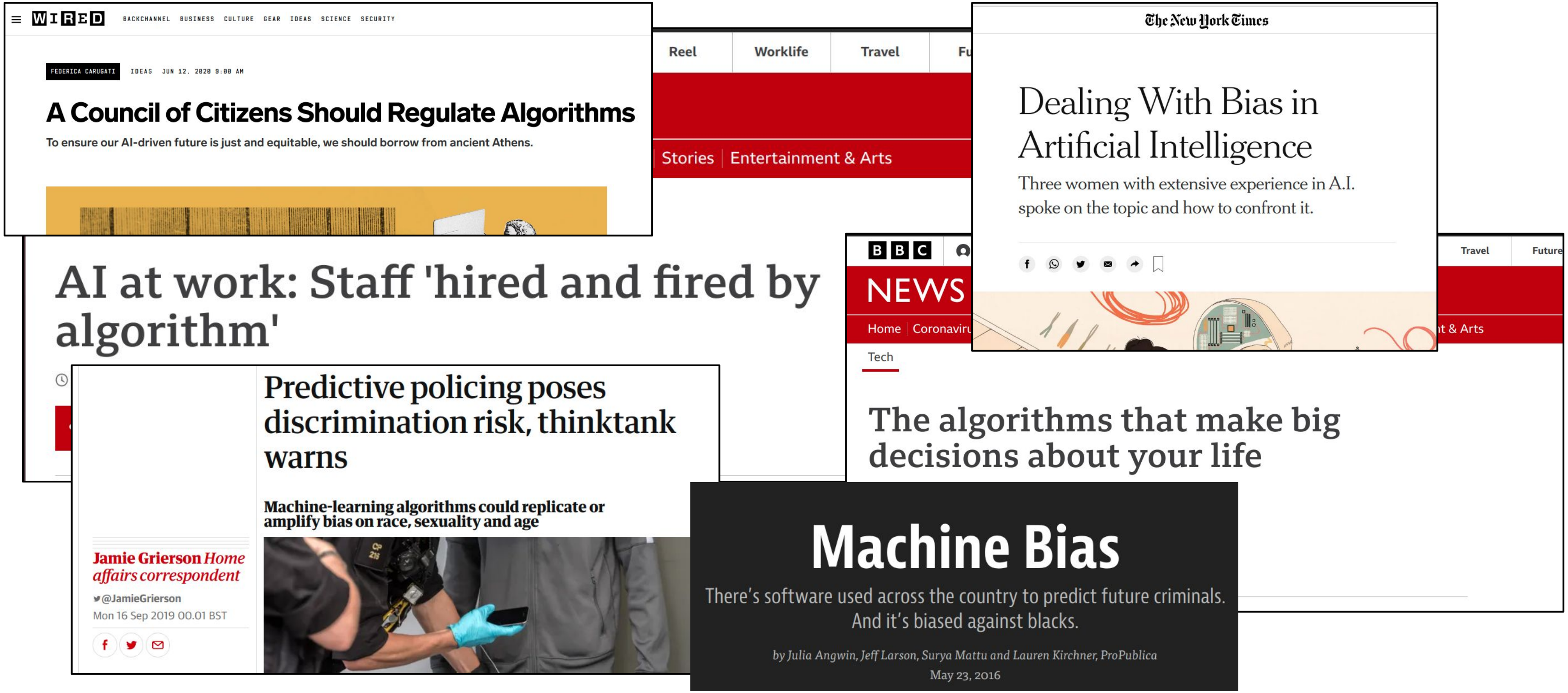
Black-Box Model

Why do we need explanations (or XAI in general)?

Automated decision-making is already being used in many scenarios:

- **Recidivism risk** [Dressel & Farid, 2018]
- **University admissions** [Waters & Miikkulainen, 2014]
- **Rejecting/Accepting a job applicant** [Liem C.C.S. et al., 2018]
- **Prescribing medications and treatments** [Yoo et al., 2019]
- …

Why do we need explanations (or XAI in general)?



WIRED
BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY
FEDERICA CARUGATI IDEAS JUN 12, 2020 9:00 AM
A Council of Citizens Should Regulate Algorithms
To ensure our AI-driven future is just and equitable, we should borrow from ancient Athens.

The New York Times
Dealing With Bias in Artificial Intelligence
Three women with extensive experience in A.I. spoke on the topic and how to confront it.

Reel Worklife Travel
Stories Entertainment & Arts

AI at work: Staff 'hired and fired by algorithm'

BBC NEWS
Home Coronaviru
Travel Future
t & Arts
Tech
The algorithms that make big decisions about your life

Predictive policing poses discrimination risk, thinktank warns
Machine-learning algorithms could replicate or amplify bias on race, sexuality and age
Jamie Grierson Home affairs correspondent
@JamieGrierson
Mon 16 Sep 2019 00.01 BST

Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Why do we need explanations (or XAI in general)?

We want to understand:

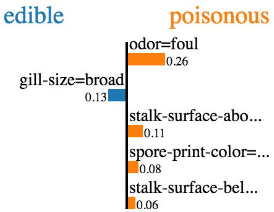
1. **Why that decision was given**
2. **How to act to obtain a desired outcome**

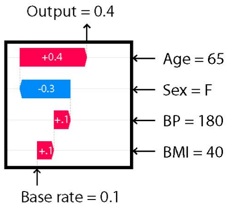
[Voigt and Von dem Bussche, 2017]

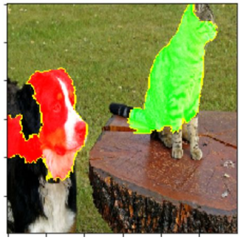
Why do we need explanations (or XAI in general)?

- **Example-based explanations**
  - Prototype and criticism [Been et al., 2016]
- **(Local/Global) Model-agnostic explanations**
  - SHAP [Lundberg and Lee, 2017]
  - LIME [Ribeiro et al., 2016]
- **Counterfactual explanations**
  - [Watcher et al., 2017]
- **Interpretable Models** (e.g., decision trees, linear models, GLM)
- **Many more!** See surveys on the topic [Adadi & Berrada, 2018]

SHAP



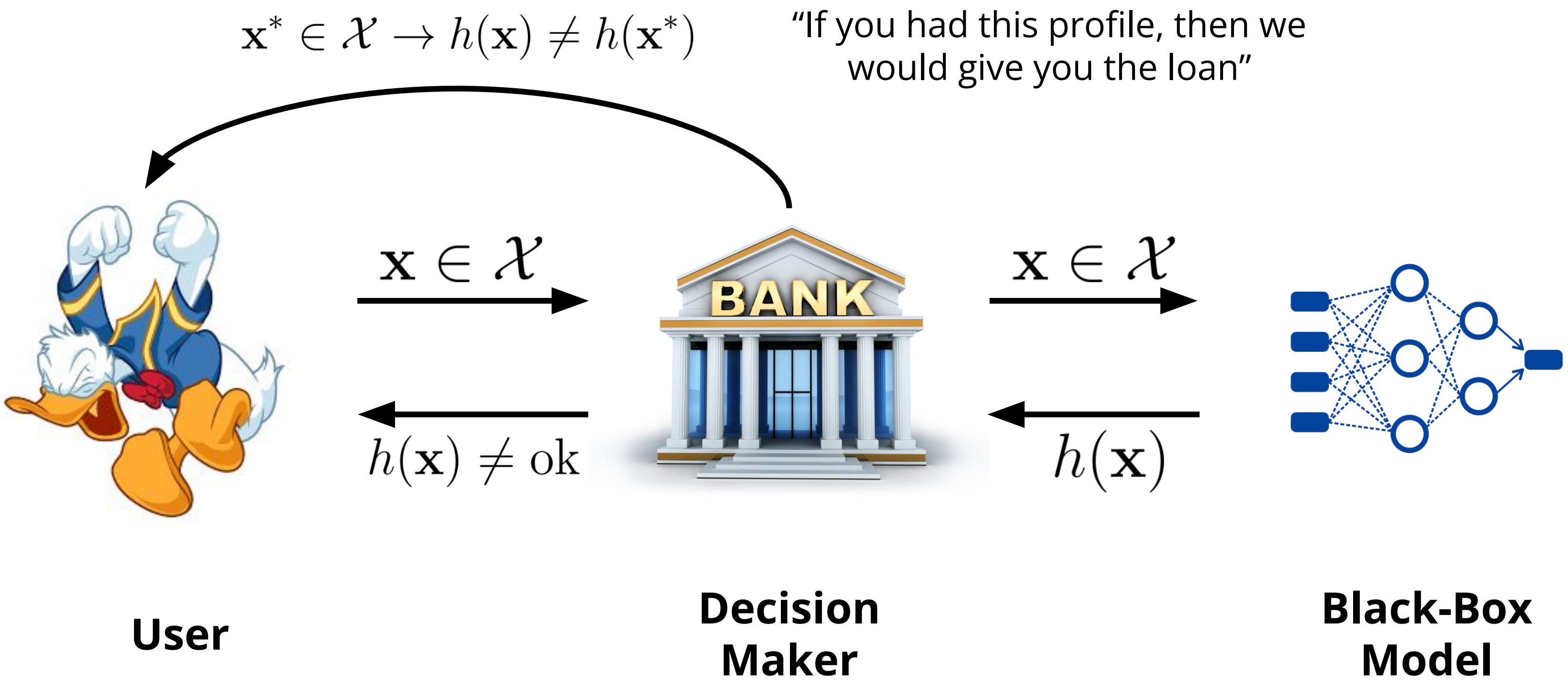




Counterfactual Explanations

A **counterfactual explanation** is a statement about “how the world would have (had) to be different for a desirable outcome to happen”

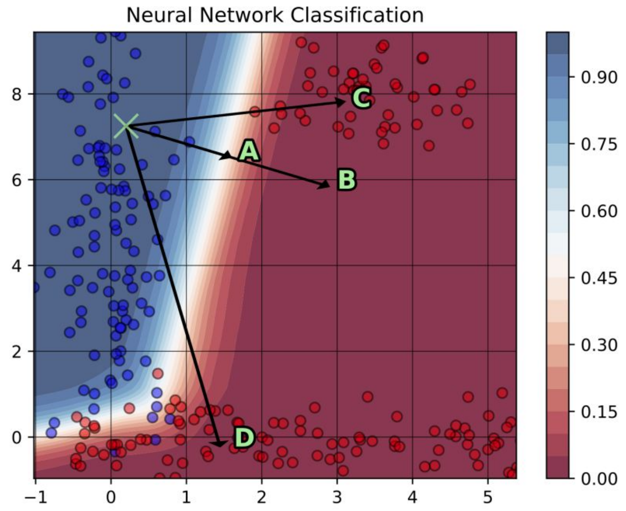
[Watcher et al., 2017; Karimi et al., 2021]


$\mathbf{x}^* \in \mathcal{X} \rightarrow h(\mathbf{x}) \neq h(\mathbf{x}^*)$
“If you had this profile, then we would give you the loan”
$\mathbf{x} \in \mathcal{X}$
$h(\mathbf{x}) \neq \text{ok}$
BANK
$\mathbf{x} \in \mathcal{X}$
$h(\mathbf{x})$
User
Decision Maker
Black-Box Model

Counterfactual Explanations

Nearest counterfactual explanations are the most similar instances of the feature vector, close to the original, that changes the prediction of the classifier.

[Watcher et al., 2017; Karimi et al., 2021]



Images taken from Poyiadzi, Rafael, et al. "FACE: feasible and actionable counterfactual explanations." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

Counterfactual Explanations

$$\mathbf{x} := \{x_0, \ldots, x_n\} \quad \mathbf{x} \in \mathcal{X}$$

$$h : \mathcal{X} \to \mathcal{Y} \quad \mathcal{Y} = \{0, 1\}$$

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

$$\mathbf{x}^* = \mathrm{argmin}_{\hat{\mathbf{x}} \in \mathcal{X}} \, d(\hat{\mathbf{x}}, \mathbf{x})$$

$$s.t.$$

$$h(\mathbf{x}) \neq h(\mathbf{x}^*)$$

$$\hat{\mathbf{x}} \in \mathcal{F}$$

[Watcher et al., 2017]

Counterfactual Explanations

- ➔ CE are **model-agnostic**
- ➔ CE **do not need** to be actual instances from the training data
- ➔ CE are **human-friendly explanations** (both **contrastive** and **selective**)
- ➔ CE are “relatively” **easy to find** (e.g., minimizing a loss function)

[Molnar, 2019]

Counterfactual Explanations [Watcher et al., 2017]

$$\mathcal{L}(\mathbf{x}, \mathbf{x}', y', \lambda) = \lambda(h(\mathbf{x}') - y') + d(\mathbf{x}, \mathbf{x}')$$

Counterfactual Explanations [Watcher et al., 2017]

$$\mathcal{L}(\mathbf{x}, \mathbf{x}', y', \lambda) = \lambda(h(\mathbf{x}') - y') + d(\mathbf{x}, \mathbf{x}')$$

$$d(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{n} \frac{|x_j - x'_j|}{MAD_j} \qquad |h(\mathbf{x}) - y'| \leq \epsilon$$

$\mathbf{x},\ y',\ \lambda$ (or ϵ) must be set in advance

Counterfactual Explanations [Watcher et al., 2017]

$$\mathcal{L}(\mathbf{x}, \mathbf{x}', y', \lambda) = \lambda(h(\mathbf{x}') - y') + d(\mathbf{x}, \mathbf{x}')$$

$$\text{argmin}_{\mathbf{x}' \in \mathcal{X}} \max_{\lambda \in \mathbb{R}} \mathcal{L}(\mathbf{x}, \mathbf{x}', y', \lambda)$$

Counterfactual Explanations

Many research works on how to build CE in the latest years:

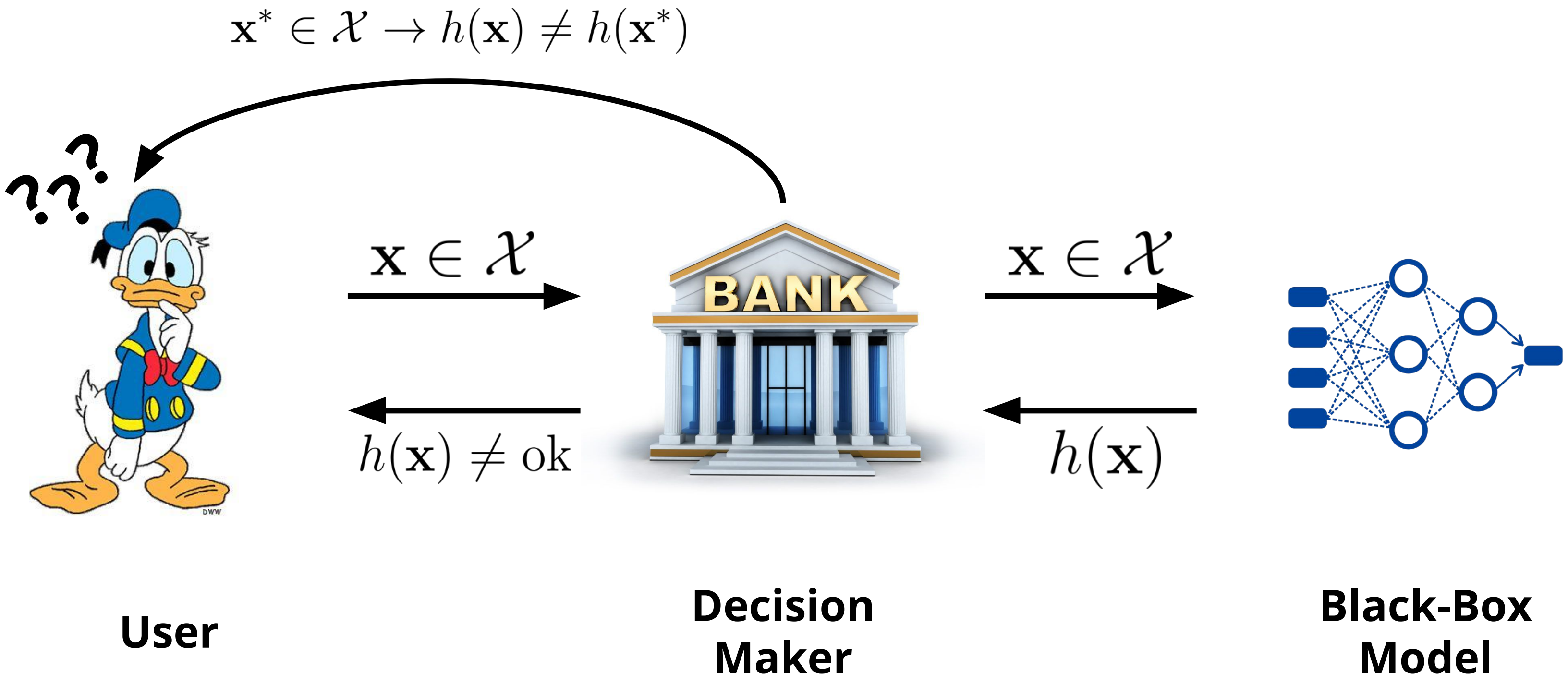
- **Multi-objective Counterfactual Explanations** [Dandl et al., 2020]
- **Counterfactual Explanations under uncertainty** [Tsirtsis et al., 2021]
- **MACE** [Karimi et al., 2020a]
- **LORE** [Guidotti et al., 2018a]
- **DICE** [Mothilal et al., 2020]
- **FACE** [Poyiadzi et al., 2020]
- ...

Many surveys on the topic (e.g., [Guidotti et al., 2018b])

Limitations of Counterfactual Explanations

- ➔ **Many CE are possible** given a single user (Rashomon Effect)
- ➔ CEs provide no **recommendations** on how to reach the given CE states
- ➔ Translating from **CEs** to **actions** is not trivial for the user
- ➔ CEs do not consider **feasibility** or the **user's effort**

[Molnar, 2019; Barocas et. al., 2020; Karimi et al., 2021; Venkatasubramanian & Alfano, 2020]


$\mathbf{x}^* \in \mathcal{X} \rightarrow h(\mathbf{x}) \neq h(\mathbf{x}^*)$
$\mathbf{x} \in \mathcal{X}$
$h(\mathbf{x}) \neq \text{ok}$
BANK
$\mathbf{x} \in \mathcal{X}$
$h(\mathbf{x})$
User
Decision Maker
Black-Box Model

Algorithmic Recourse

Algorithmic recourse is defined as “the systematic process of reversing unfavourable decisions by algorithms and bureaucracies across a range of counterfactual scenarios”

[Venkatasubramanian & Alfano, 2020; Karimi et al., 2021]

Counterfactual Interventions

➔ **Sequence of actions** instead of just a counterfactual instance

➔ They define a **cost** to mimic the **user's effort** for each action

➔ We **minimize** the cost of the sequence, given the previous constraints

➔ **Preserve qualities** of counterfactual explanations (e.g., model agnostic)

[Ustun et al., 2019; Karimi et al., 2020b; Naumann & Ntoutsi, 2021; Ramakrishnan et al., 2020]

Counterfactual Interventions

$$\mathbf{x} := \{x_0, \ldots, x_n\} \quad \mathbf{x} \in \mathcal{X}$$

$$h : \mathcal{X} \rightarrow \mathcal{Y} \quad \mathcal{Y} = \{0, 1\}$$

$$a \in \mathcal{A}$$

$$\text{cost} : \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R}^+$$

$$I^* = \text{argmin}_{I \in \mathcal{I}} \sum_{i=0}^{T} \text{cost}(a_i, \mathbf{x}_i)$$

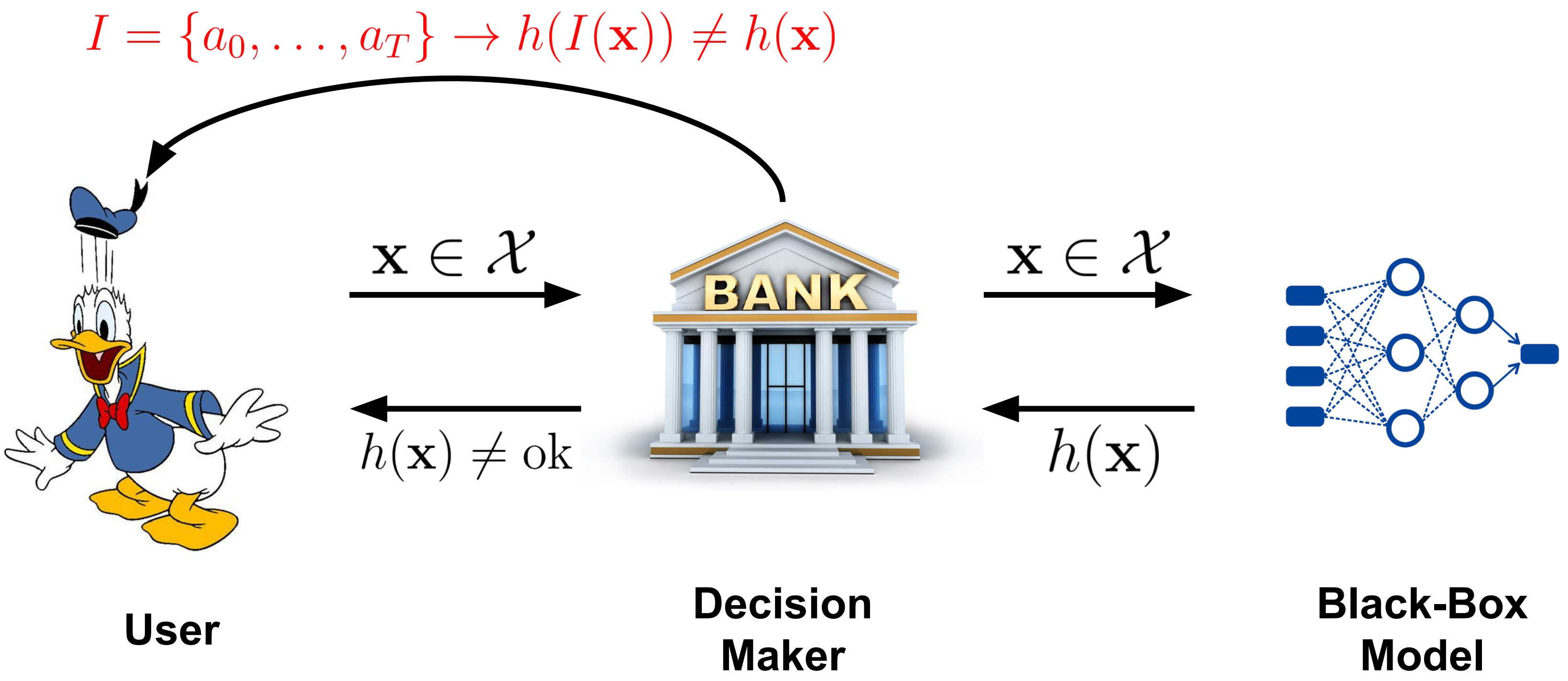
s.t.

$$I = \{a_i\}_{i=0}^{T}$$

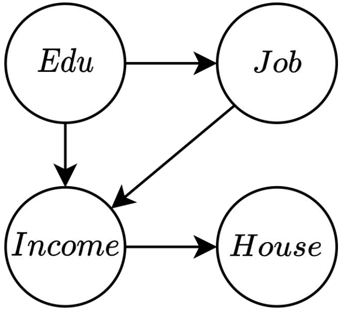
$$\mathbf{x}_t = I_{t-1}(\mathbf{x}_{t-1})$$

$$h(I(\mathbf{x_0})) \neq h(\mathbf{x_0})$$

[Ustun et al., 2019; Karimi et al., 2020b; Naumann & Ntoutsi, 2021; Ramakrishnan et al., 2020]


$I = \{a_0, \ldots, a_T\} \rightarrow h(I(\mathbf{x})) \neq h(\mathbf{x})$
$\mathbf{x} \in \mathcal{X}$
BANK
$\mathbf{x} \in \mathcal{X}$
$h(\mathbf{x}) \neq \mathrm{ok}$
$h(\mathbf{x})$
User
Decision Maker
Black-Box Model

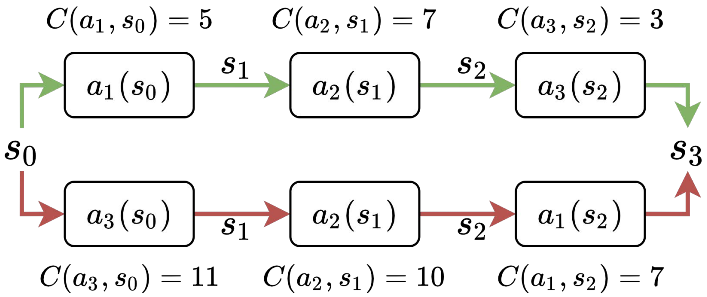
Counterfactual Interventions & Causality



a_1 : get_degree(bachelor)

a_2 : change_job(developer)

a_3 : change_house(buy)



It is **impossible** to guarantee (optimal) recourse without accessing the **true structural equations** of the causal model [Karimi et al., 2020a]

[Karimi et al., 2021; Naumann & Ntoutsi, 2021; De Toni et al., 2021]

Counterfactual Interventions

There is a growing body of research focusing of CI:

- **Recourse in linear classification** [Ustun et al., 2019]
- **SYNTH** [Ramakrishnan et al., 2020]
- **CSCF** [Naumann & Ntoutsi, 2021]
- **FastAR** [Verma et al., 2022]
- **...**

See several surveys on the topic (e.g., [Karimi et al., 2020b])

CSCF [Naumann & Ntoutsi, 2021]

$$\min_{\mathcal{S}}(\underbrace{o_1}_{\text{Sequence cost}}, \underbrace{o_2}_{\text{Gower's distance}}, \underbrace{o_{2+1}, \ldots, o_{2+h}, \ldots, o_{2+d}}_{\text{Feature tweaking frequencies}})$$

$$\text{s.t. } f(\mathbf{x}_T) = \texttt{accept} \text{ and } \bigwedge_{(a_i, v_i) \in \mathcal{S}} \mathbb{C}_i$$

Images taken from Naumann, Philip, and Eirini Ntoutsi. "Consequence-aware Sequential Counterfactual Generation." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2021.

CSCF [Naumann & Ntoutsi, 2021]

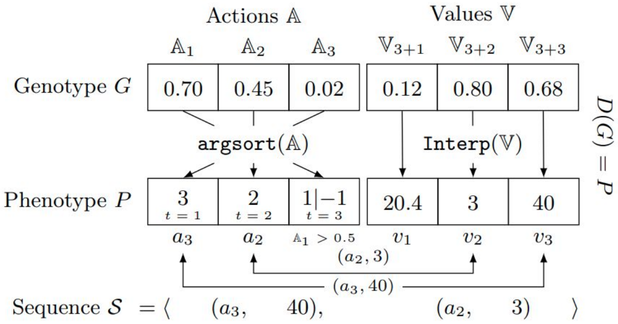
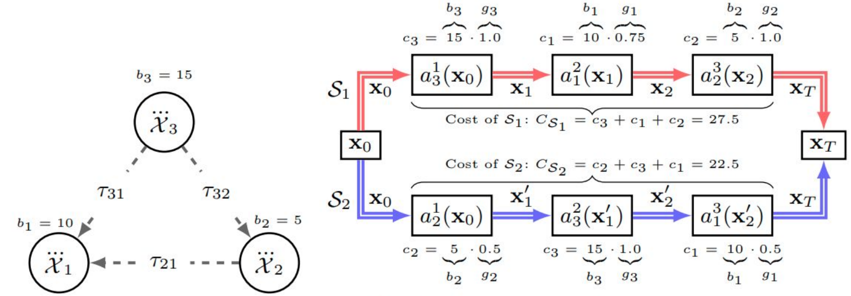


Fig. 3. Anatomy and representation of the solution decoding.



(a) Feature relationship graph $\mathcal{G}$ (b) Different sequences $\mathcal{S}_1$ (red) and $\mathcal{S}_2$ (blue)

Images taken from Naumann, Philip, and Eirini Ntoutsi. "Consequence-aware Sequential Counterfactual Generation." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2021.

CSCF [Naumann & Ntoutsi, 2021]

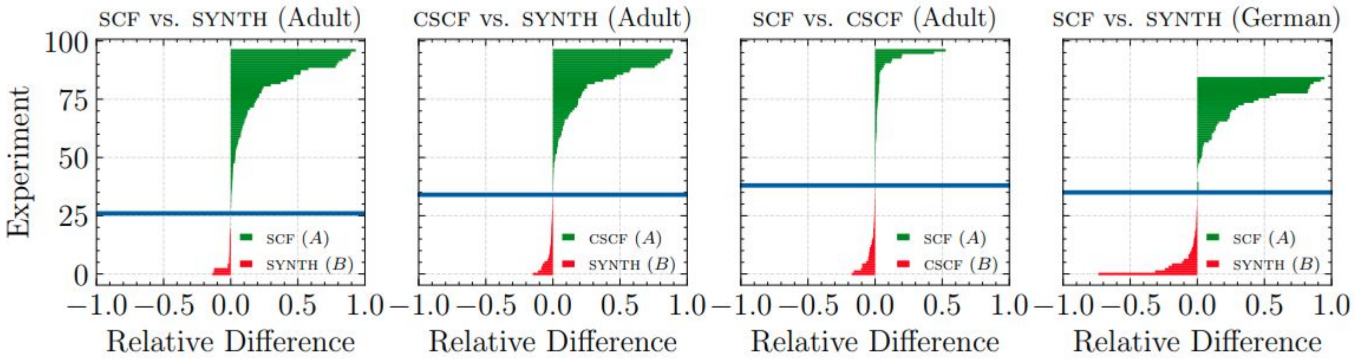


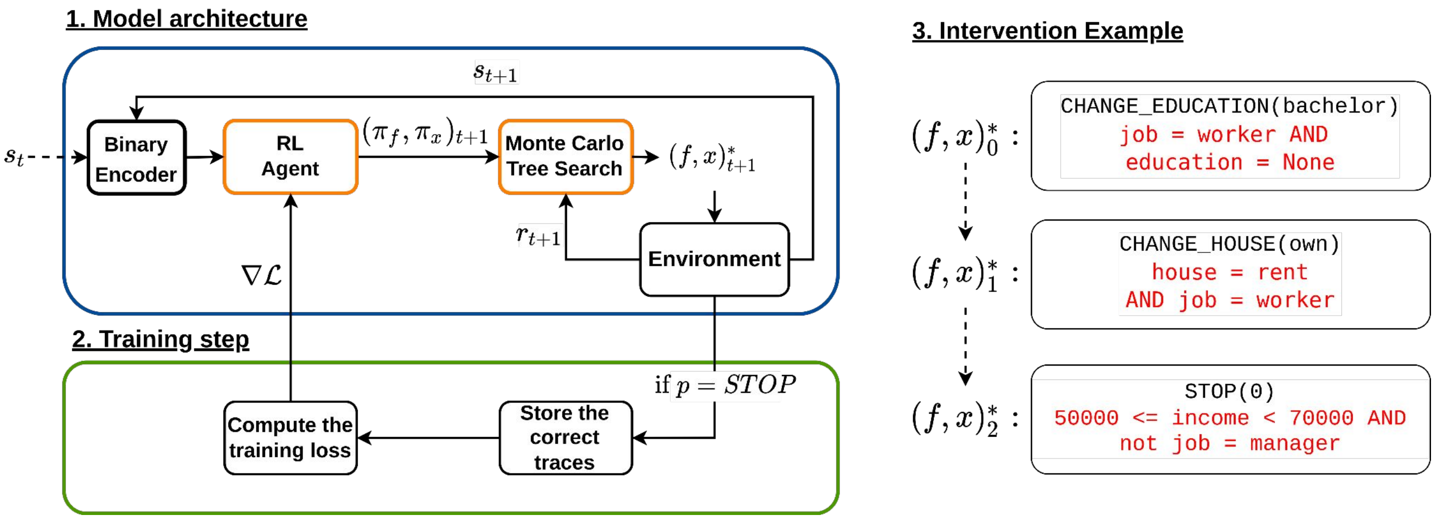
Fig. 4. Relative minimal sequence cost (o_1) differences between the three methods for both datasets and solutions with $T \leq 2$. It is computed as: $(B - A)/\max\{A, B\}$.

Images taken from Naumann, Philip, and Eirini Ntoutsi. "Consequence-aware Sequential Counterfactual Generation." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2021.

Limitations of the Counterfactual Interventions

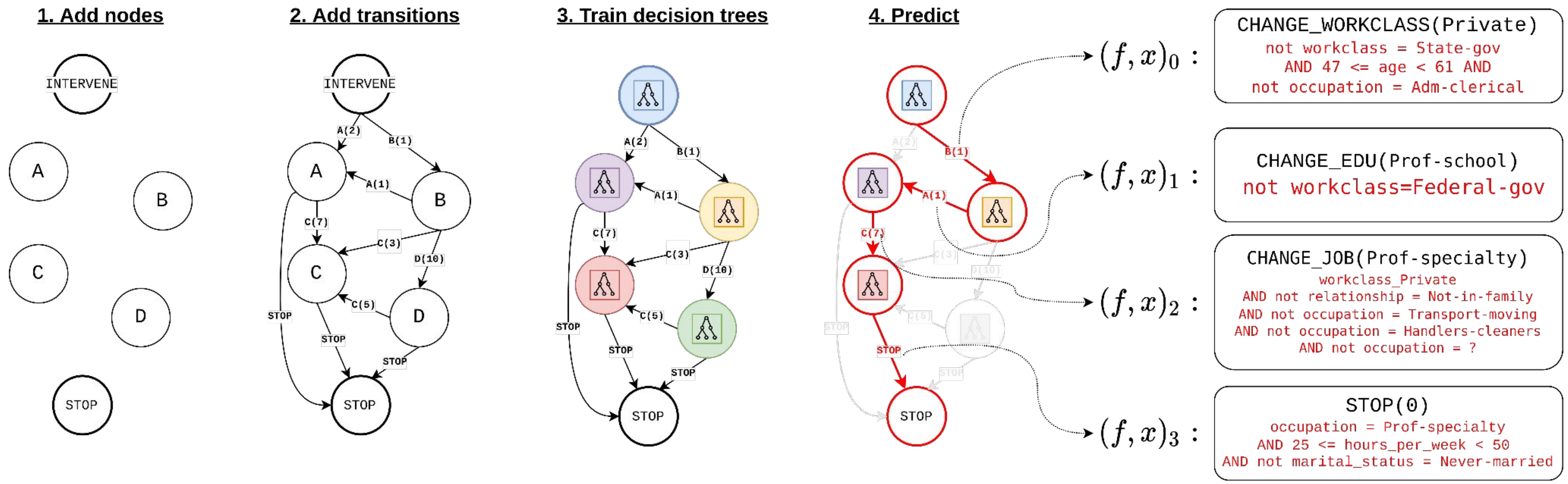
- ➔ Current methods relies on **optimization techniques**
- ➔ **Run them ex-novo** for each user (might be a **costly** process)
- ➔ Fail to explain **why** we are suggesting each intervention [Barocas et al., 2020]
- ➔ **Limitations of CFE-based recourse** [Karimi et al., 2021]

Counterfactual Interventions [De Toni et al., 2022]



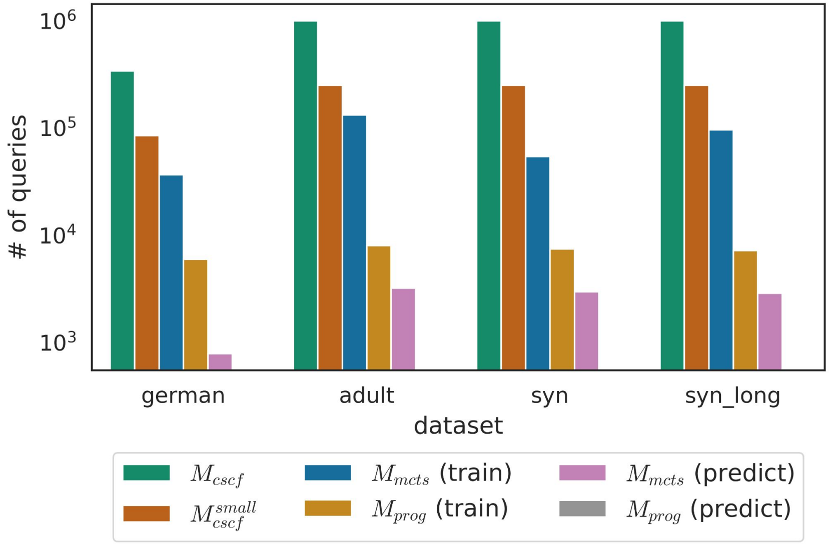
Images taken from De Toni, Giovanni, Bruno Lepri, and Andrea Passerini. "Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis." arXiv preprint arXiv:2201.07135 (2022).

Counterfactual Interventions [De Toni et al., 2022]



Images taken from De Toni, Giovanni, Bruno Lepri, and Andrea Passerini. "Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis." arXiv preprint arXiv:2201.07135 (2022).

Counterfactual Interventions [De Toni et al., 2022]



Images taken from De Toni, Giovanni, Bruno Lepri, and Andrea Passerini. "Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis." arXiv preprint arXiv:2201.07135 (2022).

Future directions

➔ Learn **costs** and the **causal graph** in a data-driven way

➔ Deal with **hidden confounders** of the causal graph

➔ **Human-in-the-loop Counterfactual Interventions**

➔ Difference between **model recommendation** and **decision**

[Barocas et al., 2020; Karimi et al., 2021; Tsirtsis & Gomez-Rodriguez, 2020]

Thank you!

SML Journal Club

https://forms.gle/XS7YqDKU9hjigR5UA

References

- **[Dressel & Farid, 2018]** Dressel, Julia, and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism." *Science advances* 4.1 (2018): eaao5580.
- **[Waters & Miikkulainen, 2014]** Waters, A., & Miikkulainen, R. (2014). GRADE: Machine Learning Support for Graduate Admissions. AI Magazine, 35(1), 64. https://doi.org/10.1609/aimag.v35i1.2504
- **[Liem et al. 2018]** Liem C.C.S. et al. (2018) Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In: Escalante H. et al. (eds) Explainable and Interpretable Models in Computer Vision and Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-319-98131-4_9
- **[Yoo et al., 2019]** Yoo, T.K., Ryu, I.H., Lee, G. et al. Adopting machine learning to automatically identify candidate patients for corneal refractive surgery. npj Digit. Med. 2, 59 (2019). https://doi.org/10.1038/s41746-019-0135-8
- **[Watcher et al., 2017]** Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harv. JL & Tech.* 31 (2017): 841.
- **[Vogit & Von dem Bussche, 2017]** Voigt, Paul, and Axel Von dem Bussche. "The eu general data protection regulation (gdpr)." *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676 (2017): 10-5555.

References

- **[Karimi et al., 2021]** Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. "Algorithmic recourse: from counterfactual explanations to interventions." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021.
- **[Molnar, 2019]** Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
- **[Dandl et al., 2020]** Dandl, Susanne, Christoph Molnar, Martin Binder, Bernd Bischl. "Multi-objective counterfactual explanations". In: Bäck T. et al. (eds) Parallel Problem Solving from Nature – PPSN XVI. PPSN 2020. Lecture Notes in Computer Science, vol 12269. Springer, Cham (2020)
- **[Guidotti et al., 2018]** Guidotti, Riccardo, et al. "Local rule-based explanations of black box decision systems." *arXiv preprint arXiv:1805.10820* (2018).
- **[Barocas et al., 2020]** Barocas, Solon, Andrew D. Selbst, and Manish Raghavan. "The hidden assumptions behind counterfactual explanations and principal reasons." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.
- **[Karimi et al., 2020a]** Karimi, A. H., von Kügelgen, L., Schölkopf, B., & Valera, I. (2020, October). Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In NeurIPS 2020.
- **[Been et al., 2016]** Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016)
- **[Adadi & Berrada, 2018]** A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
- **[Poyiadzi et al., 2020]** Poyiadzi, Rafael, et al. "FACE: feasible and actionable counterfactual explanations." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.

References

- **[Venkatasubramanian & Alfano, 2020]** Venkatasubramanian, Suresh, and Mark Alfano. "The philosophical basis of algorithmic recourse." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.
- **[Karimi et al., 2020b]** Karimi, Amir-Hossein, et al. "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects." *arXiv preprint arXiv:2010.04050* (2020).
- **[Guidotti et al., 2018b]** Guidotti, Riccardo, et al. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51.5 (2018): 1-42.
- **[Ustun et al., 2019]** Ustun, Berk, Alexander Spangher, and Yang Liu. "Actionable recourse in linear classification." *Proceedings of the conference on fairness, accountability, and transparency*. 2019.
- **[Naumann & Ntoutsi, 2021]** Naumann, Philip, and Eirini Ntoutsi. "Consequence-aware Sequential Counterfactual Generation." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2021.
- **[Ramakrishnan et al., 2020]** Ramakrishnan, Goutham, Yun Chan Lee, and Aws Albarghouthi. "Synthesizing action sequences for modifying model decisions." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 04. 2020.
- **[De Toni et al., 2022]** De Toni, Giovanni, Bruno Lepri, and Andrea Passerini. "Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis." arXiv preprint arXiv:2201.07135 (2022).
- **[Tsirtsis & Gomez-Rodriguez, 2020]** Tsirtsis, Stratis, and Manuel Gomez Rodriguez. "Decisions, counterfactual explanations and strategic behavior." *Advances in Neural Information Processing Systems* 33 (2020): 16749-16760.
- **[Lundberg & Lee, 2017]** Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” Advances in Neural Information Processing Systems (2017).
- **[Ribeiro et al., 2016]** 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. ACM (2016)

Acknowledgements

- Donald Duck - https://www.adesivipareti.com/25445-thickbox/adesivi-e-vinile-donald-duck-italian-6366.jpg
- Donald Duck angry - https://mystickermania.com/sticker-packs/disney-cartoons/angry-donald-duck
- Donald Duck confused - https://www.pinterest.com/pin/512143788866854143/
- Bank - https://www.adnyfinance.com/836.html
- Neural Network - https://thenounproject.com/icon/neural-network-3339036/