Seminar on **AI Safety**

Kickoff Meeting

Holger H. Hoos[1,2,3] Marie Anastacio[1] Jakob Bossek[1]

[1]Dept. of Computer Science, RWTH Aachen University, Germany
[2]LIACS, Universiteit Leiden, The Netherlands
[3]University of British Columbia, Vancouver, BC, Canada

21 October 2022

RWTH AACHEN UNIVERSITY

Neural networks are great

Neural networks are great

but ...

Neural networks are great

but ... sensitive to input perturbations:

horn

hot dog

Source: https://kennysong.github.io/adversarial.js/

Neural networks are great

but ... sensitive to input perturbations:

Stop

120 km/h

Source: https://kennysong.github.io/adversarial.js/

Neural networks are great

but ... sensitive to input perturbations:

Stop

120 km/h

Source: https://kennysong.github.io/adversarial.js/

$\rightsquigarrow$ lack of robustness, vulnerability to adversarial attacks

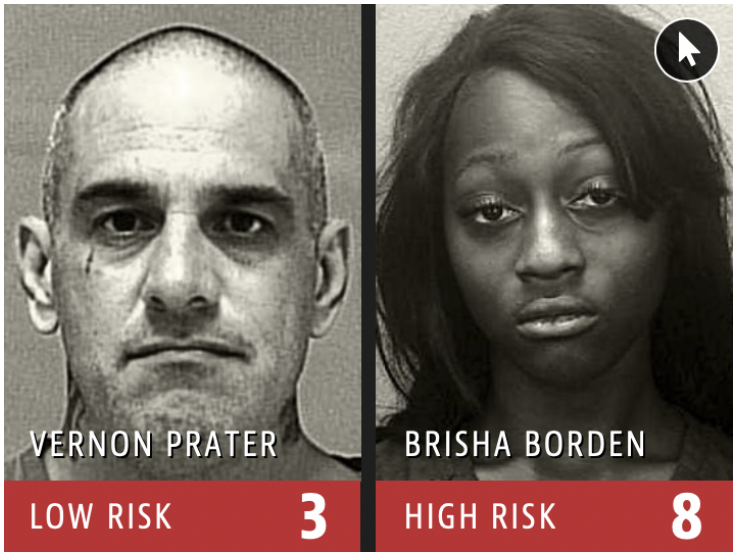
Machine learning models can be very powerful

Machine learning models can be very powerful

but ...

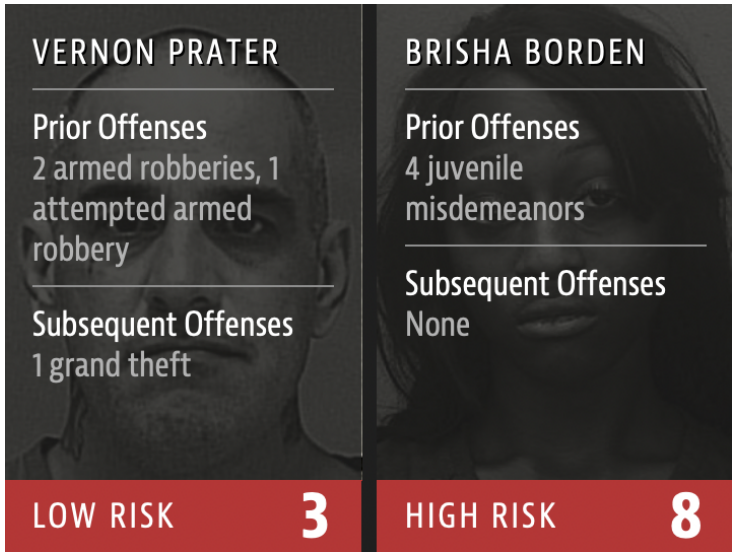
Machine learning models can be very powerful
but ... may be biased:

Machine learning models can be very powerful
but ... may be biased:



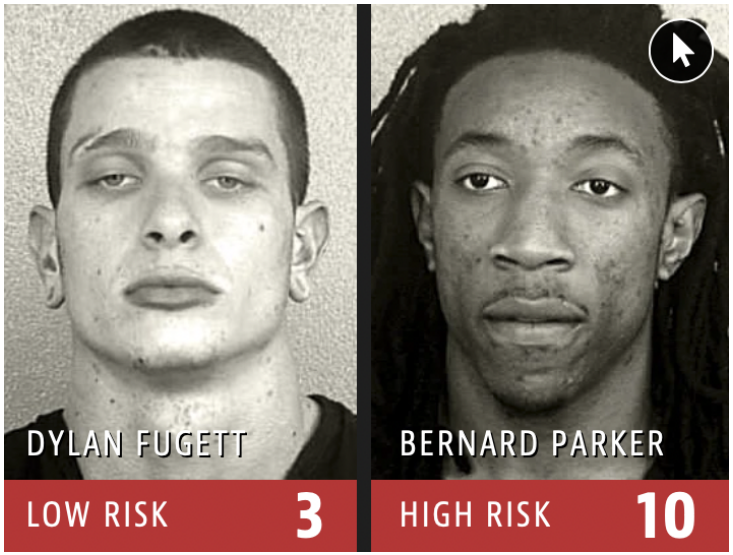
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

Machine learning models can be very powerful
but ... may be biased:



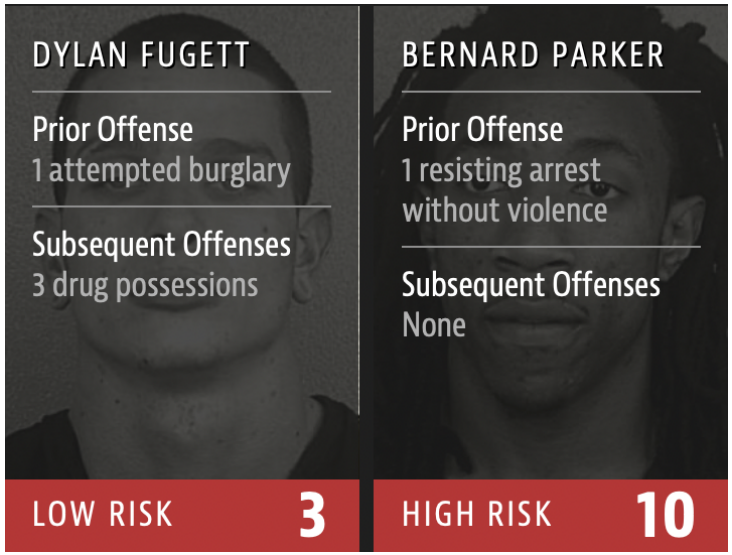
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

Machine learning models can be very powerful
but ... may be biased:



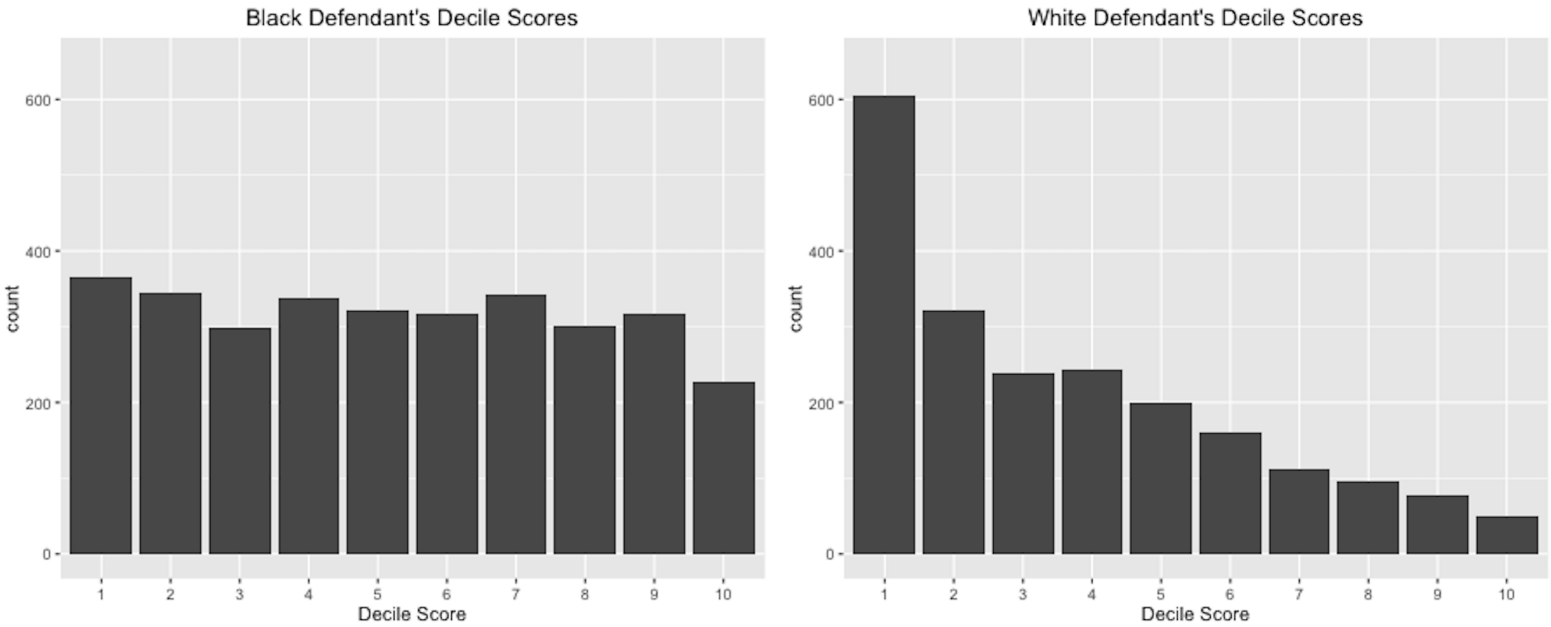
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

Machine learning models can be very powerful
but ... may be biased:



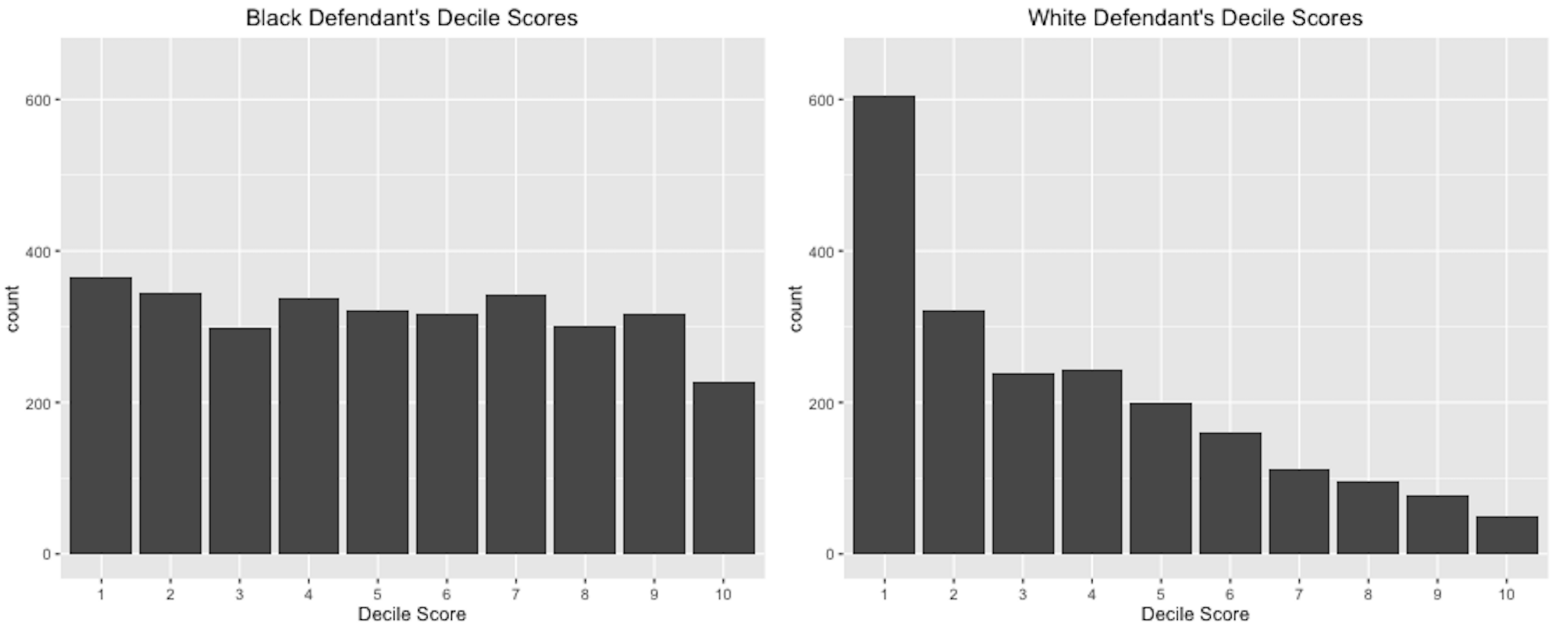
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

Machine learning models can be very powerful
but ... may be biased:



Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

Machine learning models can be very powerful
but ... may be biased:



Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

$\rightsquigarrow$ e.g., lack of trustworthiness, amplification of bias in the real-world

Machine learning models may be explainable

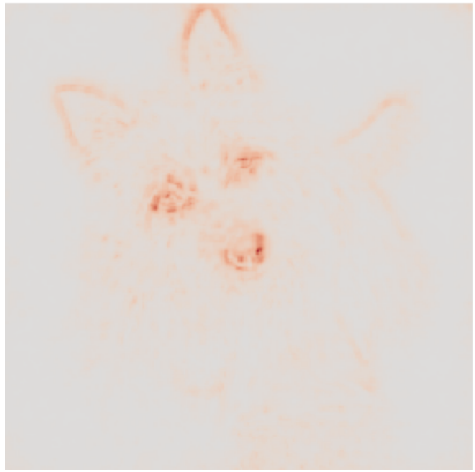
Machine learning models may be explainable

(a) Husky classified as wolf

(b) Explanation

Source: "Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro *et. al.*, ACM SIGKDD 2016.

Machine learning models may be explainable
but ...explanations are also vulnerable



Source: Explanations can be manipulated and geometry is to blame

Ann-Kathrin Dombrowski *et. al.*, NeurIPS 2019.

Machine learning models may be explainable
but ...explanations are also vulnerable

Source: Explanations can be manipulated and geometry is to blame
Ann-Kathrin Dombrowski *et. al.*, NeurIPS 2019.

Prof. Dr. Holger H. Hoos

Alexander von Humboldt Professor
Chair for AI Methodology (AIM)
Department of Computer Science
RWTH Aachen University

E-Mail: hh@aim.rwth-aachen.de
Website: https://hoos.ca/

Research interests

- Intersection of machine learning, automated reasoning and optimisation
- Automated design and analysis of algorithms: performance prediction, algorithm configuration, algorithm selection and construction of parallel algorithm portfolios
- Iterated Local Search (ILS) algorithms
- Bio-inspired optimisation, in particular Ant Colony Optimization (ACO)
- Bioinformatics and computer music

M. Sc. Marie Anastacio

Postdoctoral researcher
Chair for AI Methodology (AIM)
Department of Computer Science
RWTH Aachen University

E-Mail: anastacio@aim.rwth-aachen.de

Research interests

- Algorithm configuration
- Automated Artificial Intelligence
- Heuristic Optimisation
- Model-based Optimisation
- Robustness and explainability

Dr. Jakob Bossek

Assistant Professor (Akademischer Rat)
Chair for AI Methodology (AIM)
Department of Computer Science
RWTH Aachen University

E-Mail: bossek@aim.rwth-aachen.de
Website: http://www.jakobbossek.de/

Research interests

- Heuristic Optimisation (in particular Evolutionary Algorithms)
- Combinatorial (Multi-Objective) Optimisation
- Evolutionary Diversity Optimisation (EDO) and Quality Diversity (QD)
- Theory of randomised search heuristics
- Sequential Model-Based Optimisation (SMBO)
- Instance Generation for Benchmarking (in particular for the TSP)
- Algorithm Selection and Configuration

Important dates (take note!)

- Progress update (via e-mail, bullet points are OK, but do give us some details): 18 November 2022, 18:00 CEST (hard deadline!)
- Final report due (PDF via e-mail): 27 January 2023, 18:00 CEST (hard deadline!)

Groups and topics I

RV-1 Konstantin Geisler, Rene Heinz-Peter Evertz

Topic: Adversarial attacks

Christian Szegedy et al. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014

RV-2 Luk Jonas Fuchs, Nick Valentin Kocher

Topic: Formal verification

Vincent Tjeng, Kai Yuanqing Xiao, and Russ Tedrake. "Evaluating Robustness of Neural Networks with Mixed Integer Programming". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019

Groups and topics II

RV-3 Kiana Mishelle Adamik, Nicolas Schumann

Topic: Heuristic verification

Nicholas Carlini and David A. Wagner. "Towards Evaluating the Robustness of Neural Networks". In: *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 39–57. DOI: 10.1109/SP.2017.49

EI-1 Nhu-Yen Nguyen, Marius Andre Jean-Michel Peterfalvi

Topic: Outcome explanation (global)

Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. ed. by Isabelle Guyon et al. 2017, pp. 4765–4774

Groups and topics III

EI-2 Torge Schöwing, Jan Brinkmann

Topic: Outcome explanation (local)

Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319

EI-3 Ozan Ege Sap, Peter Benjamin Schwarz

Topic: Inspection

Julia Moosbauer et al. "Explaining Hyperparameter Optimization via Partial Dependence Plots". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato et al. 2021, pp. 2280–2291

Groups and topics IV

EI-4 Dobromir Iordanov Panayotov, Katharina Kösseler

Topic: Counterfactual explanations

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. Ed. by Mireille Hildebrandt et al. ACM, 2020, pp. 607–617. DOI: 10.1145/3351095.3372850

Groups and topics V

P-1 Yitong Guo, Niklas Fücker

Topic: Model inversion

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*. Ed. by Indrajit Ray, Ninghui Li, and Christopher Kruegel. ACM, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677

Groups and topics VI

P-2 Phillip Ahlers, Mohamed-Wassim Deghdagh

Topic: Inference

Karan Ganju et al. "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*. Ed. by David Lie et al. ACM, 2018, pp. 619–633. DOI: 10.1145/3243734.3243834

Groups and topics VII

B-1 Archit Dhama, Marco Bischoff

Topic: Mitigating bias via sampling strategies

F Kamiran and TGK Calders. "Classification with no discrimination by preferential sampling". In: *Informal proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'10, Leuven, Belgium, May 27-28, 2010)*. 2010, pp. 1–6

Groups and topics VIII

B-2 Xiyang Yang, Julian Treiber

Topic: Mitigating bias via data alteration

Hao Wang, Berk Ustun, and Flávio P. Calmon. "Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6618–6627

Take-home messages

- AI systems need to be robust and trustworthy
- This seminar will cover a wide range of AI safety methods and challenges
- We're here to help – do not hesitate to contact us if you have questions