

Odpowiedzialna Sztuczna Inteligencja



Przemysław Biecek
Samsung R&D Institute Poland

15/1/2021

Principal Data Scientist w **Samsung R&D Institute Poland** (AI Team),
ex Netezza, IBM, Disney.

Profesor uczelni na Politechnice Warszawskiej, wydział Matematyki
i Nauk Informacyjnych. Założyciel zespołu **MI2DataLab**.



**eXplainable
Artificial
Intelligence
Working
Group**

Obszar zainteresowań: Interakcja człowiek-model,
Uczenie maszynowe, Zastosowania modeli predykcyjnych
w medycynie i finansach



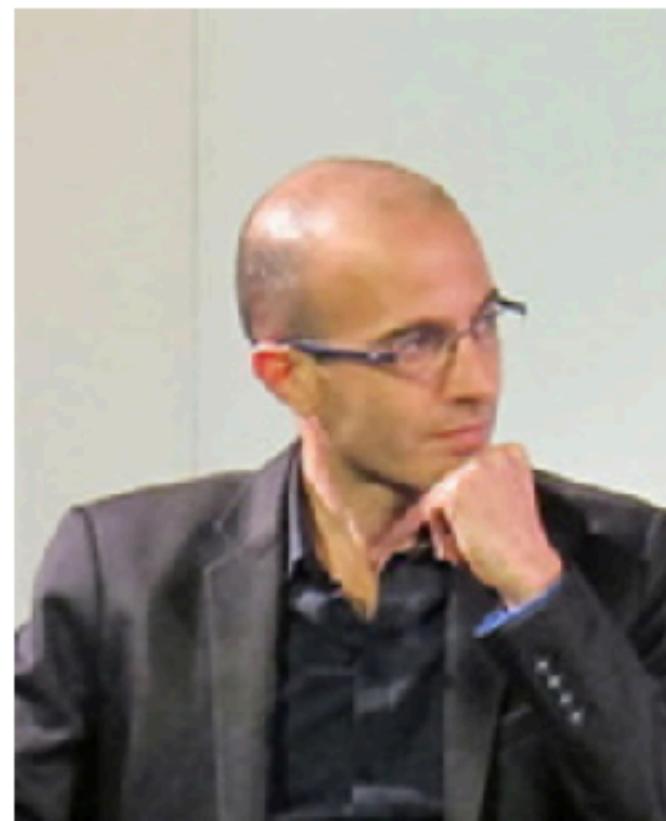
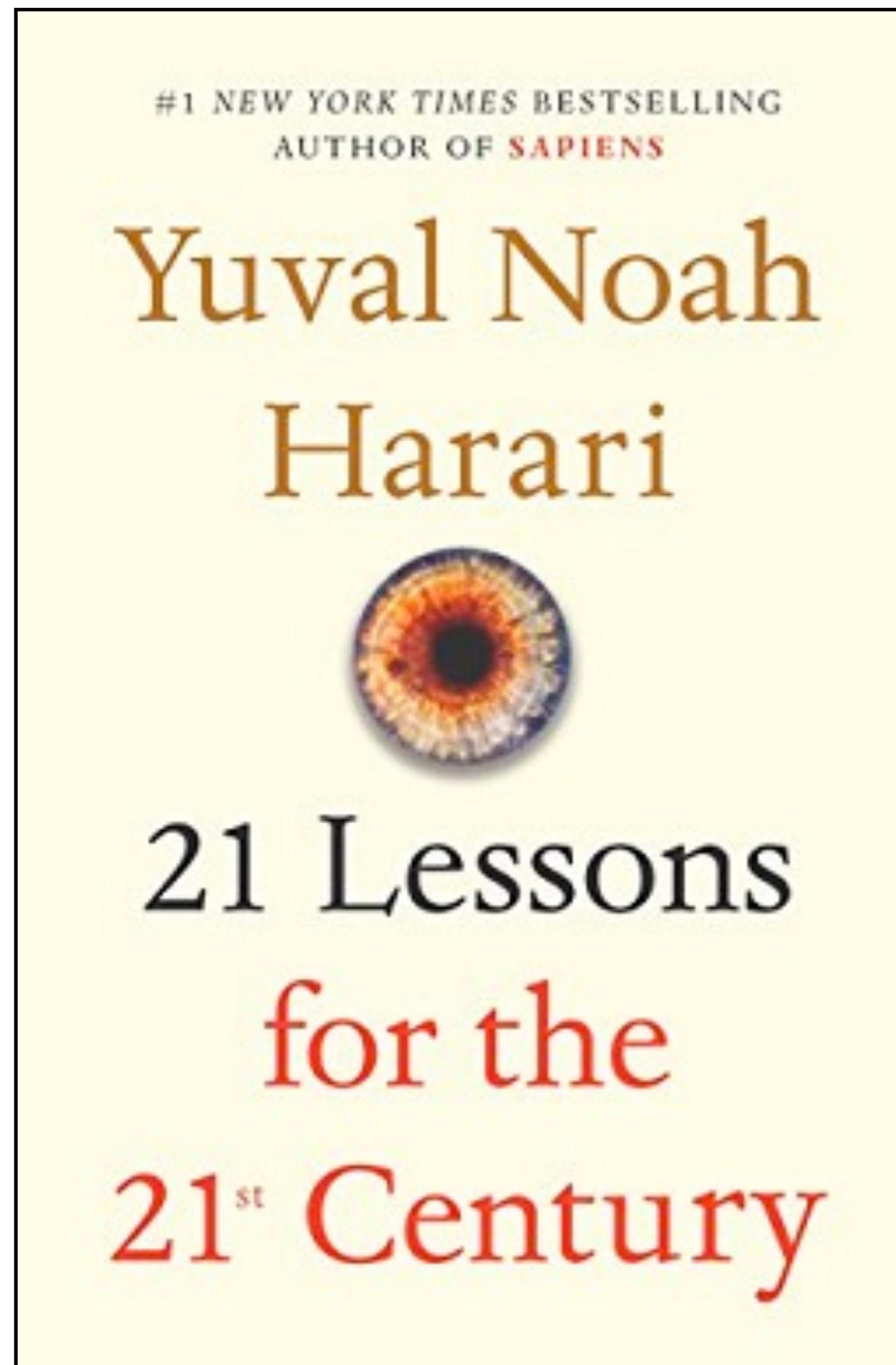
Disclaimer:

Celem tej prezentacji jest przedstawienie obecnego stanu wiedzy oraz trendów związanych z tematyką Responsible AI.

Ta prezentacja nie potwierdza, ani nie zaprzecza, że firma Samsung stosuje lub rekomenduje zaprezentowane metody.

O czym jest ta prezentacja?

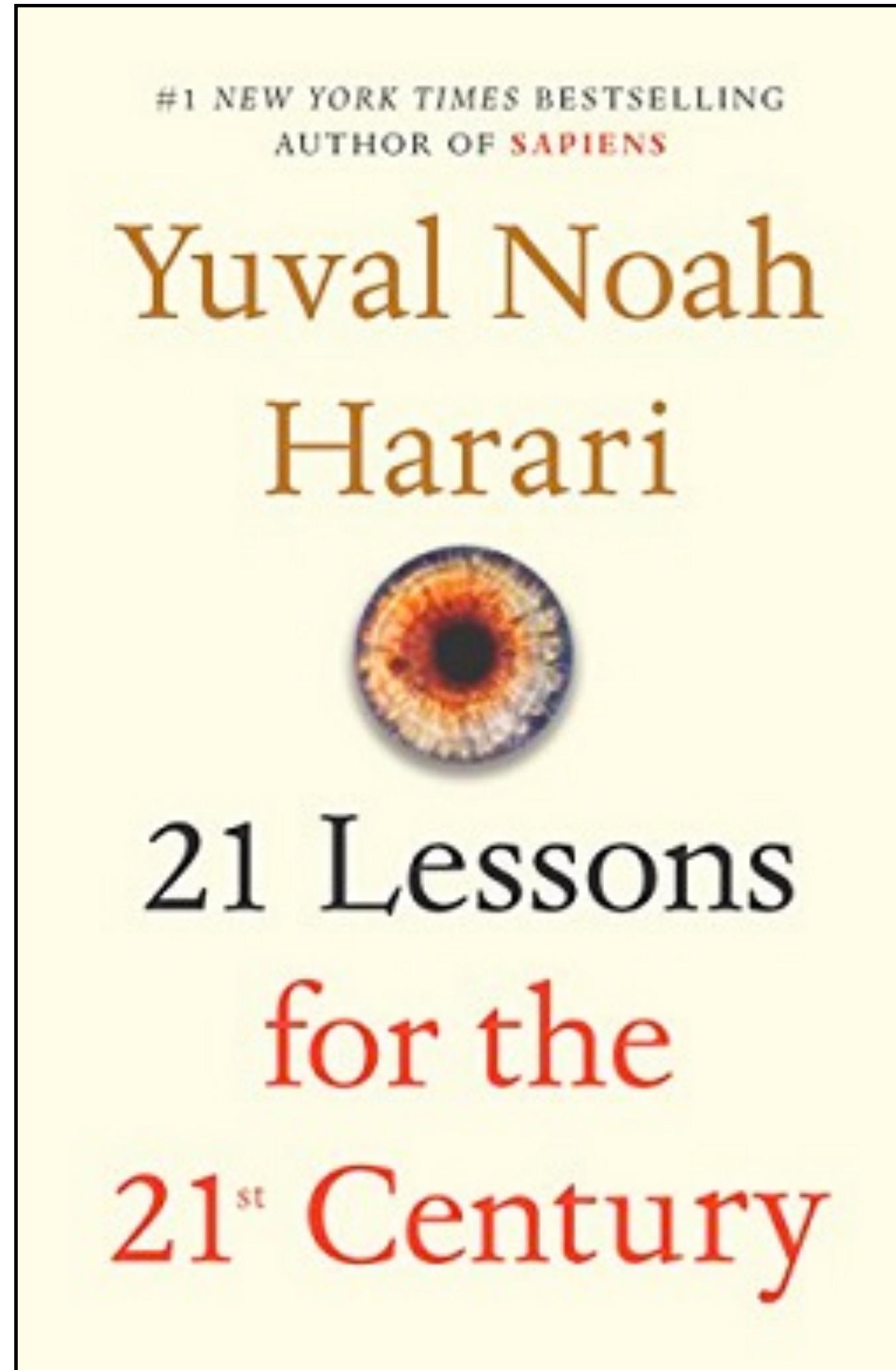
O czym jest ta prezentacja?



“Indeed, many movies about artificial intelligence are so divorced from scientific reality that one suspects they are just allegories of completely different concerns. Thus the 2015 movie Ex Machina seems to be about an AI expert who falls in love with a female robot only to be duped and manipulated by her. But in reality, this is not a movie about the human fear of intelligent robots. It is a movie about the male fear of intelligent women, and in particular the fear that female liberation might lead to female domination. Whenever you see a movie about an AI in which the AI is female and the scientist is male, it’s probably a movie about feminism rather than cybernetics. For why on earth would an AI have a sexual or a gender identity? Sex is a characteristic of organic multicellular beings. What can it possibly mean for a non-organic cybernetic being?”

— Yuval Noah Harari, *21 Lessons for the 21st Century*

O czym jest ta prezentacja?



AI to zbiór algorytmów wspierających proces decyzyjny.

Mniej i bardziej złożonych algorytmów jest wokół nas wiele

- high frequency trading,
- scoring kredytowy,
- kwalifikacja do szczepień,
- rekomendacja terapii (np. operacji chirurgicznej),
- real time bidding dla reklam,
- spersonalizowane filtry w programie pocztowym.



Piotr Gliński

@PiotrGlinski

...

Replying to [@PiotrGlinski](#)

O tym, kto dostał wsparcie, nie decydowały sympatie, rodzaj uprawianej sztuki, ale algorytm pokazujący kto stracił przychody w wyniku pandemii. To nie są pieniądze dla jednej osoby, ale dla całych zespołów ludzi, którzy z dnia na dzień stracili środki do życia. 2/2

[Translate Tweet](#)

12:44 PM · Nov 14, 2020 from Warsaw, Poland · Twitter for iPhone

61 Retweets 87 Quote Tweets 152 Likes



Ministerstwo Kultury doprecyzowuje kryteria oceny wniosków

W pierwotnym komunikacie MKiDN na ten temat zabrakło informacji o algorytmie, o którym wspomniał wicepremier. Dopiero w drugiej informacji, zamieszczonej w sobotę 14 listopada, pada, że wysokość rekompensat była obliczana na podstawie danych księgowych i statystycznych za rok 2019. Poszczególne kryteria dotyczyły:

- spadku przychodów instytucji/przedsiębiorcy 2020 vs 2019
- liczby zatrudnionych osób
- liczby odwołanych imprez/wydarzeń artystycznych
- wpływu dofinansowania na lokalną społeczność
- udziału przychodów z muzyki tańca i teatru w ogólnych przychodach (w odniesieniu do przedsiębiorców)
- zadłużenia przedsiębiorców spowodowanego pandemią.

O czym jest ta prezentacja?

AI to zbiór algorytmów wspierających proces decyzyjny.

Algorytmy nie są opracowywane wyłącznie na bazie wiedzy dziedzinowej.

Algorytmy AI budowane są tak by **automatycznie** uwzględniały zależności występujące w (często bardzo **dużych**) korpusach **danych**.

Czy to istotny temat?

Czy to istotny temat?



Czy to istotny temat?



Moim zdaniem:

AI jest jak
energia jądrowa

Czy to istotny temat?

- Bardzo zaawansowana i wyspecjalizowana technologia, której rozwój wymaga kapitału ludzkiego, aparaturowego i wysokiej jakości danych.
- Może być użyta zarówno jako broń jak i jako lekarstwo.
- Rodzi wiele obaw, tak uzasadnionych jak i nieuzasadnionych.

Moim zdaniem:
**AI jest jak
energia jądrowa**

Czy to istotny temat?

Algorytmy AI pozwalają na zachowanie urządzeń w sposób bardziej zgodny z naszymi intencjami i oczekiwaniami.



W czym problem?

Podstawowe założenie modeli
uczenia maszynowego:

Przyszłość będzie podobna do
przeszłości

Podstawowe założenie modeli

uczenia maszynowego:

COVID19

Przyszłość będzie podobna do

przeszłości

Lista porażek w zastosowaniach AI

Więcej na: <https://romanlutz.github.io/ResponsibleAI/>

Speech Detection

- Oh dear... AI models used to flag hate speech online are, er, racist against black people
- The Risk of Racial Bias in Hate Speech Detection
- Toxicity and Tone Are Not The Same Thing: analyzing the new Google API on toxicity, PerspectiveAPI.
- Voice Is the Next Big Platform, Unless You Have an Accent
- Google's speech recognition has a gender bias
- Fair Speech report by Stanford Computational Policy Lab, also covered in [Speech recognition algorithms may also have racial bias](#)
- Automated moderation tool from Google rates People of Color and gays as "toxic"
- Someone made an AI that predicted gender from email addresses, usernames. It went about as well as expected

Image Labelling & Face Recognition

- Google Photos identified two black people as 'gorillas'
- When It Comes to Gorillas, Google Photos Remains Blind
- The viral selfie app ImageNet Roulette seemed fun – until it called me a racist slur
- Google Is Investigating Why it Trained Facial Recognition on 'Dark Skinned' Homeless People
- Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
- Machines Taught by Photos Learn a Sexist View of Women
- Tenants sounded the alarm on facial recognition in their buildings. Lawmakers are listening.
- Google apologizes after its Vision AI produced racist results

Public Benefits & Health

- A health care algorithm affecting millions is biased against black patients
- What happens when an algorithm cuts your health care
- China Knows How to Take Away Your Health Insurance
- Foretelling the Future: A Critical Perspective on the Use of Predictive Analytics in Child Welfare

Lending & Credit approval

- Gender Bias Complaints against Apple Card Signal a Dark Side to Fintech
- Exploring Racial Discrimination in Mortgage Lending: A Call for Greater Transparency
- DFS Issues Guidance to Life Insurers on Use of "External Data" in Underwriting Decisions

Hiring

- Amazon scraps secret AI recruiting tool that showed bias against women
- Automated Employment Discrimination
- Help wanted: an examination of hiring algorithms, equity, and bias
- All the Ways Hiring Algorithms Can Introduce Bias
- Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices
- Help Wanted – An Examination of Hiring Algorithms, Equity, and Bias
- Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.

Employee evaluation

- Houston Schools Must Face Teacher Evaluation Lawsuit
- How Amazon automatically tracks and fires warehouse workers for 'productivity'

Pre-trial risk assessment and criminal sentencing

- Machine Bias
- How We Analyzed the COMPAS Recidivism Algorithm
- GitHub repository for COMPAS analysis



OECD AI Principles overview

The OECD AI Principles promote use of AI that is innovative and trustworthy and that respects human rights and democratic values. Adopted in May 2019, they set standards for AI that are practical and flexible enough to stand the test of time.

Values-based principles



Inclusive growth, sustainable development and well-being >



Human-centred values and fairness >



Transparency and explainability >



Robustness, security and safety >



Accountability >

Recommendations for policy makers



Investing in AI research and development >



Fostering a digital ecosystem for AI >



Shaping an enabling policy environment for AI >



Building human capacity and preparing for labour market transformation >



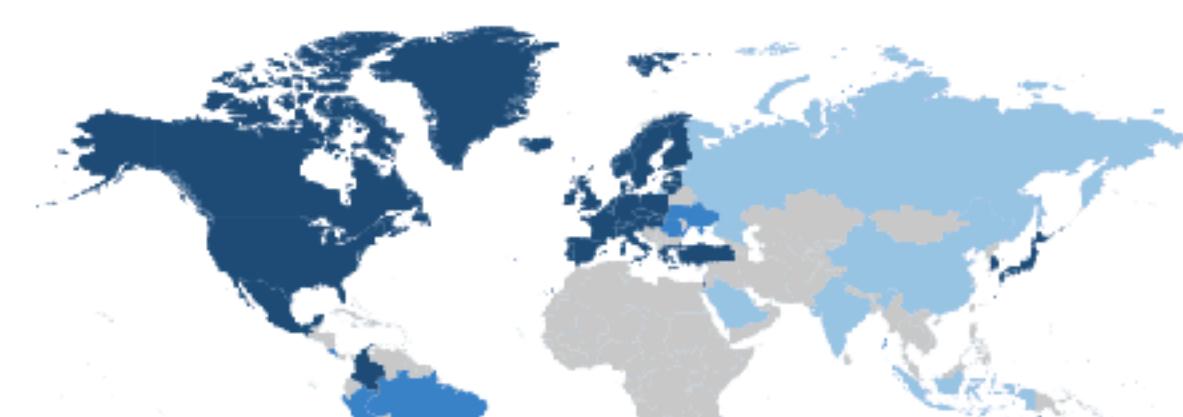
International co-operation for trustworthy AI >

The OECD AI Principles focus on how governments and other actors can shape a human-centric approach to trustworthy AI. As an OECD legal instrument, the principles represent a common aspiration for its adhering countries.



» Official text

Governments that have committed to the AI Principles



<https://oecd.ai/ai-principles>

Principles for AI Ethics



1. Fairness

The company will strive to apply the values of **equality** and **diversity** in AI system throughout its entire lifecycle.

The company will strive not to reinforce nor propagate negative or unfair bias.

The company will strive to provide **easy access** to all users.

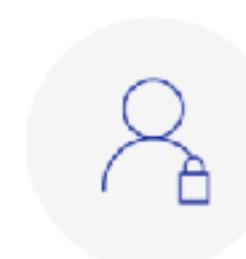


2. Transparency

Users will be aware that they are interacting with AI.

AI will be **explainable** for users to understand its decision or recommendation to the extent technologically feasible.

The process of collecting or utilizing personal data will be **transparent**.



3. Accountability

The company will strive to apply the principles of **social and ethical responsibility** to AI.

AI system will be **adequately protected** and have **security measures** to prevent data bre

The company will strive to **benefit the society** and promote the **corporate citizenship** th

Trusted AI Lifecycle through Open Source

Pillars of trust, woven into the lifecycle of an AI application



Did anyone tamper with it?



Is it fair?



Is it easy to understand?



Is it accountable?



Is it lineage?

The screenshot shows the TensorFlow website with a navigation bar including 'Install', 'Learn', 'API', 'Resources', 'Community', and 'Why TensorFlow'. The 'Resources' dropdown is open, showing 'Responsible AI' as the selected item. Below the navigation, there are links for 'Models & datasets', 'Tools', 'Libraries & extensions', 'TensorFlow Certificate program', and 'Learn ML'. The main content area features a section titled 'What is Responsible AI?' with a sub-section about the development of AI creating opportunities and raising questions about building AI systems that benefit everyone. It includes links to 'Adversarial Robustness 360', 'AI Fairness 360', 'AI Explainability 360', and 'In the works!'. A search bar and a star icon are also present.

The screenshot shows a dark-themed article from McKinsey & Company. The title is 'Leading your organization to responsible AI'. The article discusses the impact of AI increasing across sectors and societies, emphasizing the need for fairness and interpretability. It includes sections on 'Recommended best practices for AI', 'Fairness', and 'Interpretability'. At the bottom, there is a link to 'Learn more about Google's Responsible AI Practices' and a date of 'May 2, 2019 | Article'.

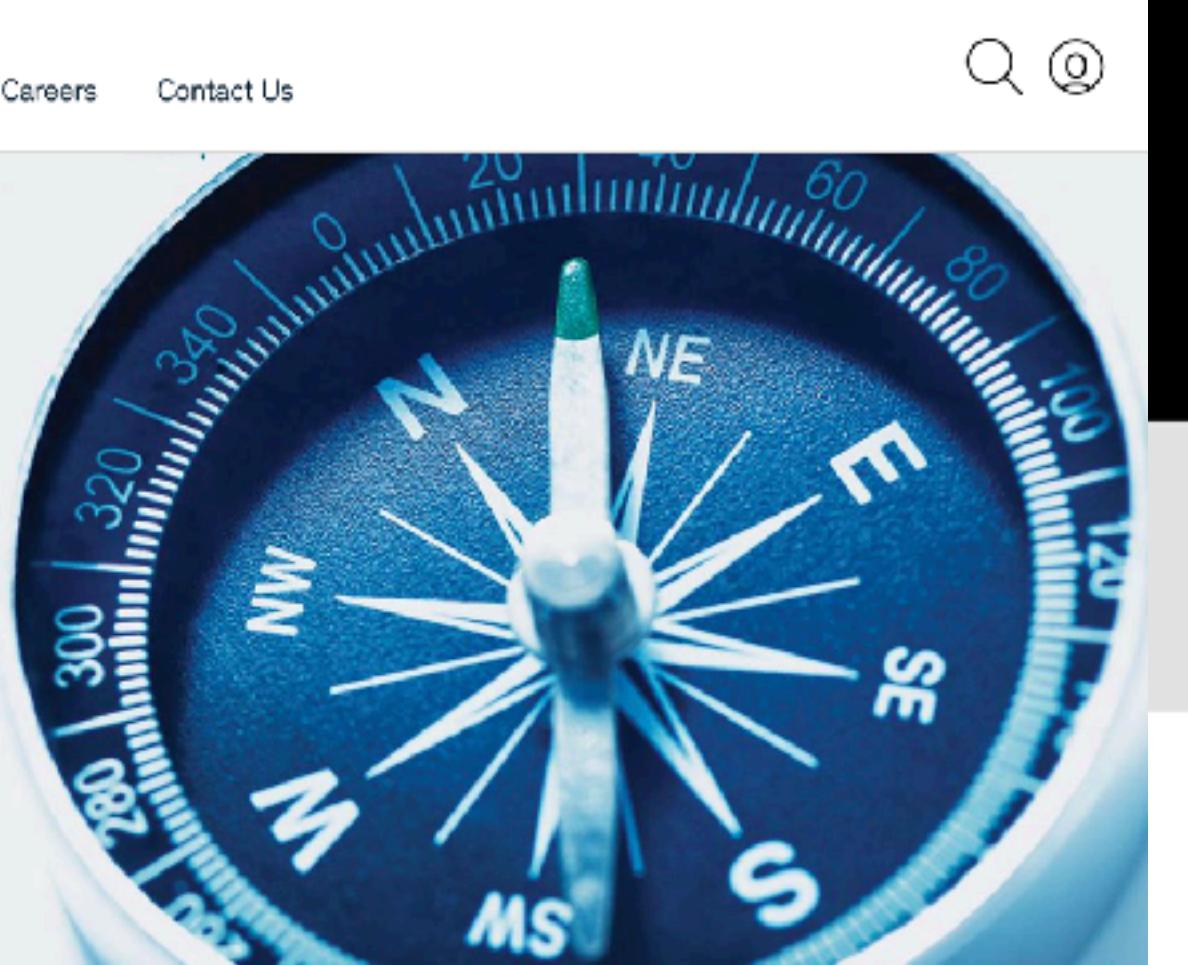


Overview

What is Responsible AI?

Explainable AI and the pursuit of using technology and statistical methods to explain Machine Learning models, quickly became a much larger question. Best practices in applying AI is not just a statistical question, but a people and process question as well, which forms the key elements of Responsible AI. In order to achieve maximum transparency and understanding of AI, it is imperative to address and understand the full view of models and their Impact. There are six categories that comprise the most critical themes in

Machine Learning AI, and



A practical guide to
Responsible Artificial
Intelligence (AI)



pwc.com/rai

Principles for AI Ethics

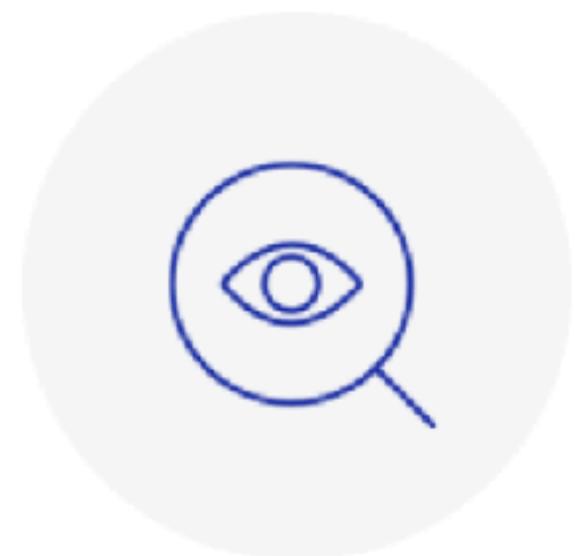


1. Fairness

The company will strive to apply the values of **equality and diversity** in AI system throughout its entire lifecycle.

The company will strive not to **reinforce nor propagate negative or unfair bias**.

The company will strive to provide **easy access** to all users.



2. Transparency

Users will be aware that they are **interacting with AI**.

AI will be explainable for users to understand its decision or recommendation to the extent technologically feasible.

The process of collecting or utilizing personal data will be **transparent**.



3. Accountability

The company will strive to apply the principles of **social and ethical responsibility** to AI system

AI system will be **adequately protected** and have **security measures** to prevent data breach and cyber attacks.

The company will strive to **benefit the society** and promote the **corporate citizenship** though AI system.

Fairness

Równość traktowania

The Apple Card credit lines



Steve Wozniak @stevewoz · Nov 10

Replying to [@dhh](#) and [@AppleCard](#)

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

91

573

3.4K



Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist



RESEARCH & KNOWLEDGE



DIGITAL

SUSTAINABILITY

LEADERSHIP

COMPETITIVENESS

[Research & Knowledge](#) > [Articles](#) > Amazon's sexist hiring algorithm could still be better than a human

5 min.

November 2018

Related topics:

Digital

Talent Management

Disruption

Strategy

Amazon's sexist hiring algorithm could still be better than a human

Expecting algorithms to perform perfectly might be asking too much of ourselves

Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist



RESEARCH & KNOWLEDGE

DIGITAL

SUSTAINABILITY

LEADERSHIP

COMPETITIVENESS

[Research & Knowledge](#) > [Articles](#) > Amazon's sexist hiring algorithm could still be better than a human

5 min.

November 2018

Related topics:

Digital

Talent Management

Disruption

Strategy

Amazon's sexist hiring algorithm could still be better than a human

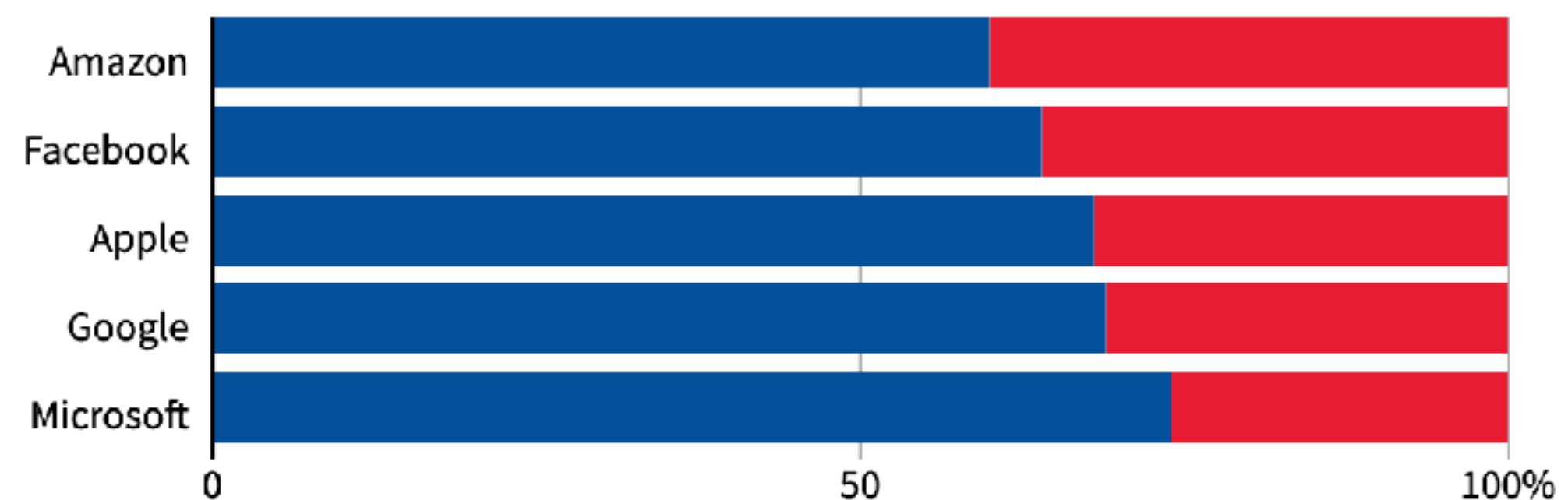
Expecting algorithms to perform perfectly might be asking too much of ourselves

Dominated by men

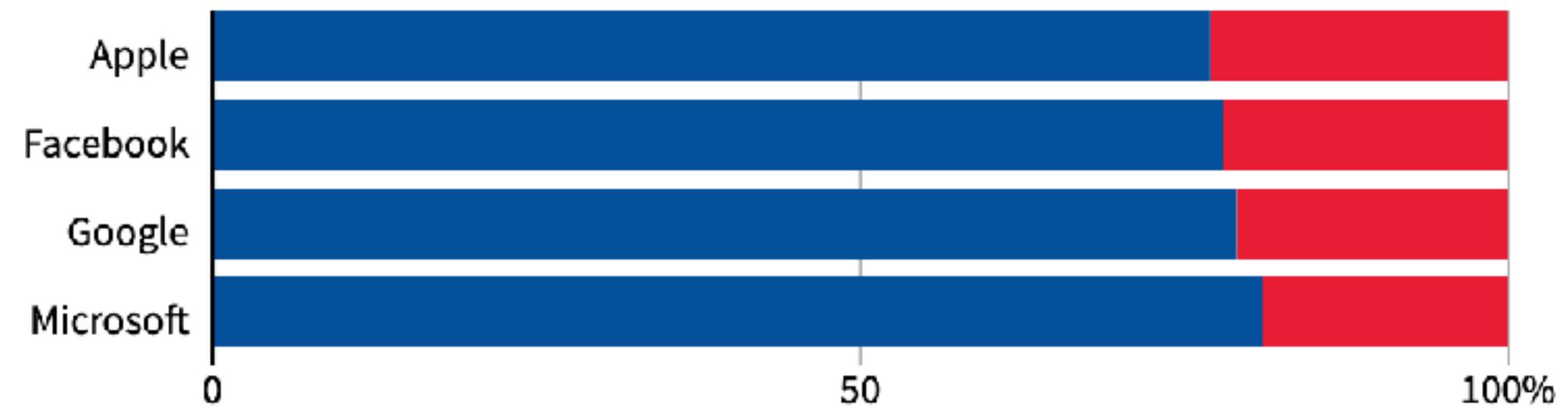
Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

GLOBAL HEADCOUNT

■ Male ■ Female



EMPLOYEES IN TECHNICAL ROLES



Note: Amazon does not disclose the gender breakdown of its technical workforce.

COMPAS Recidivism Risk Score Data and Analysis

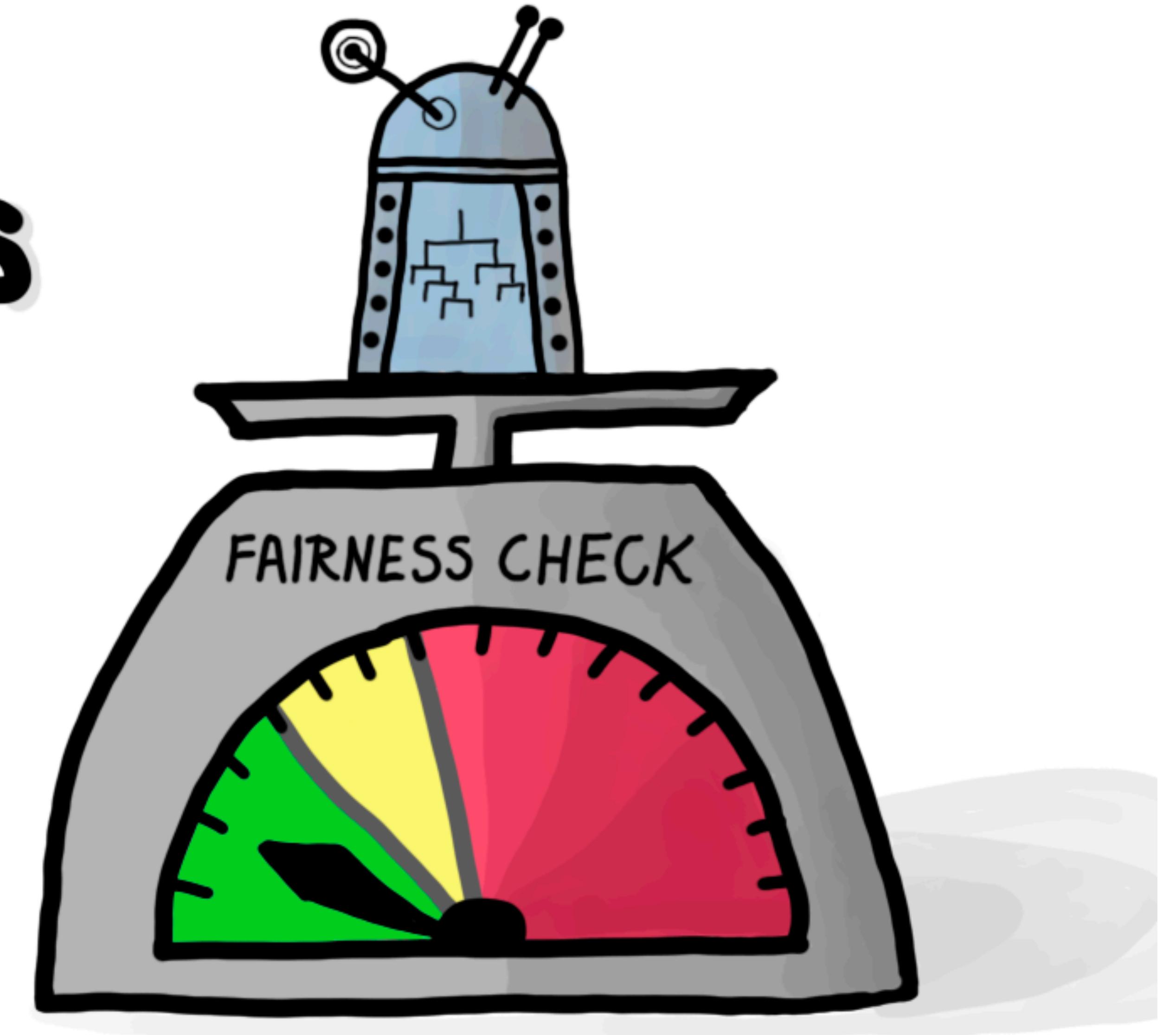
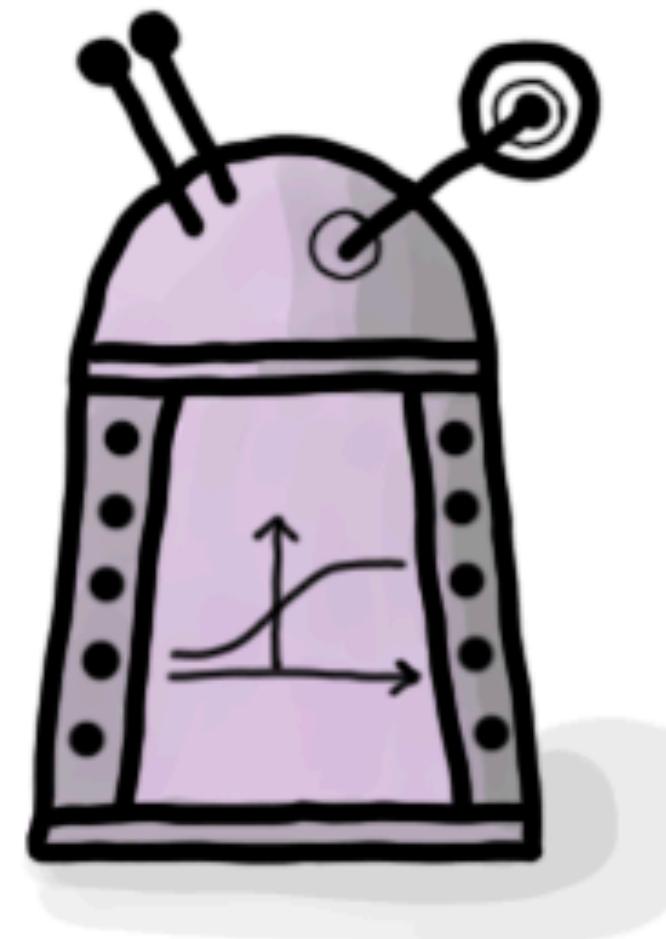
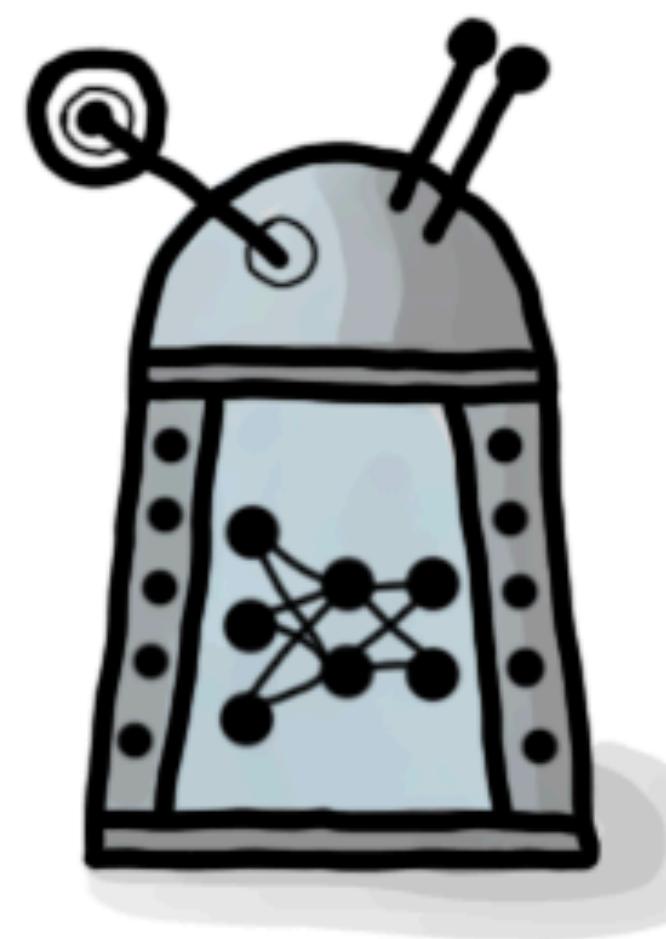


Fairness in Machine Learning

Limitations and Opportunities

What about applications that aren't about people? Consider "Street Bump," a project by the city of Boston to crowdsource data on potholes. The smartphone app automatically detects pot holes using data from the smartphone's sensors and sends the data to the city. Infrastructure seems like a comfortably boring application of data-driven decision-making, far removed from the ethical quandaries we've been discussing.

fairmodels

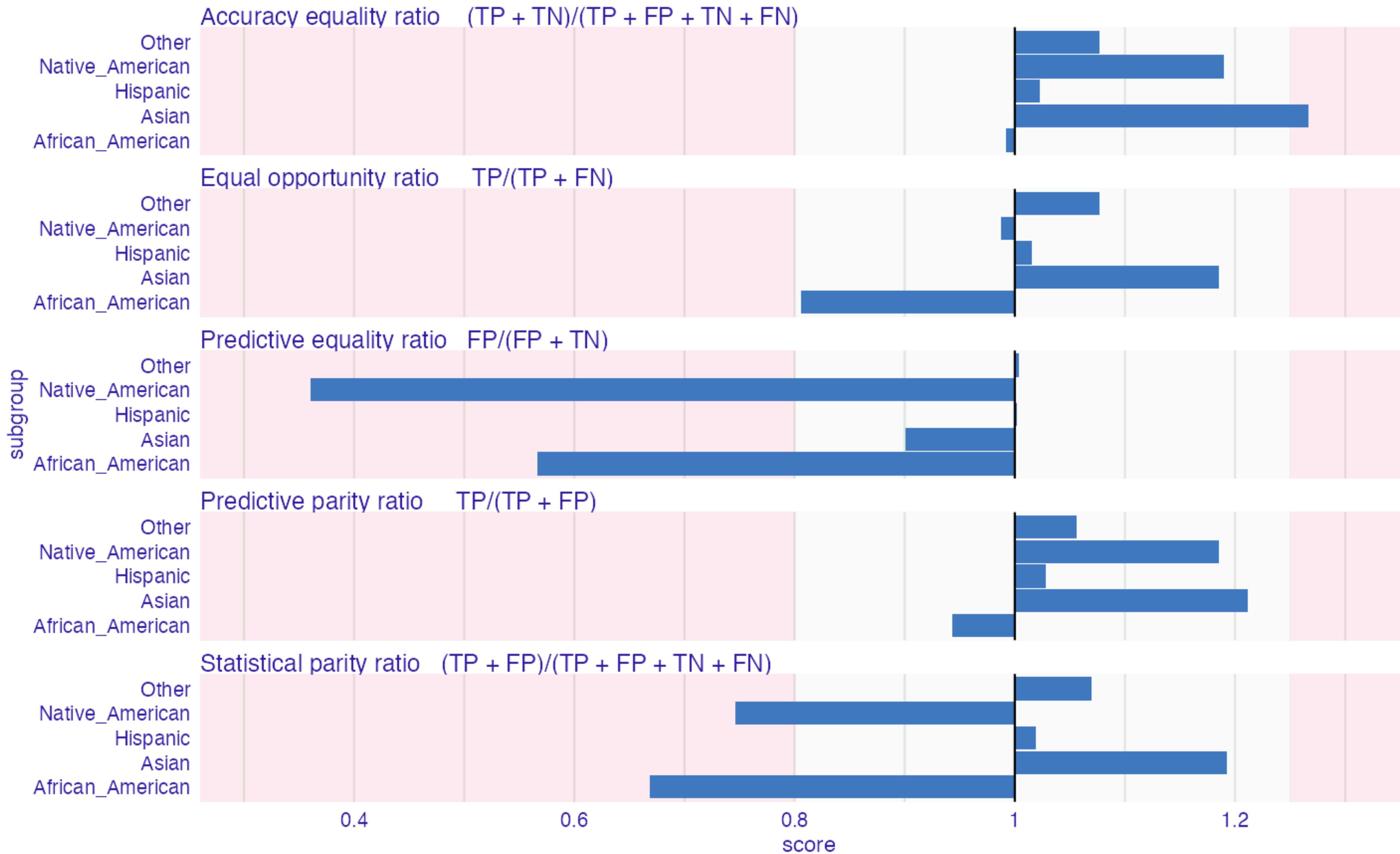


<https://modeloriented.github.io/fairmodels/>

Fairness check

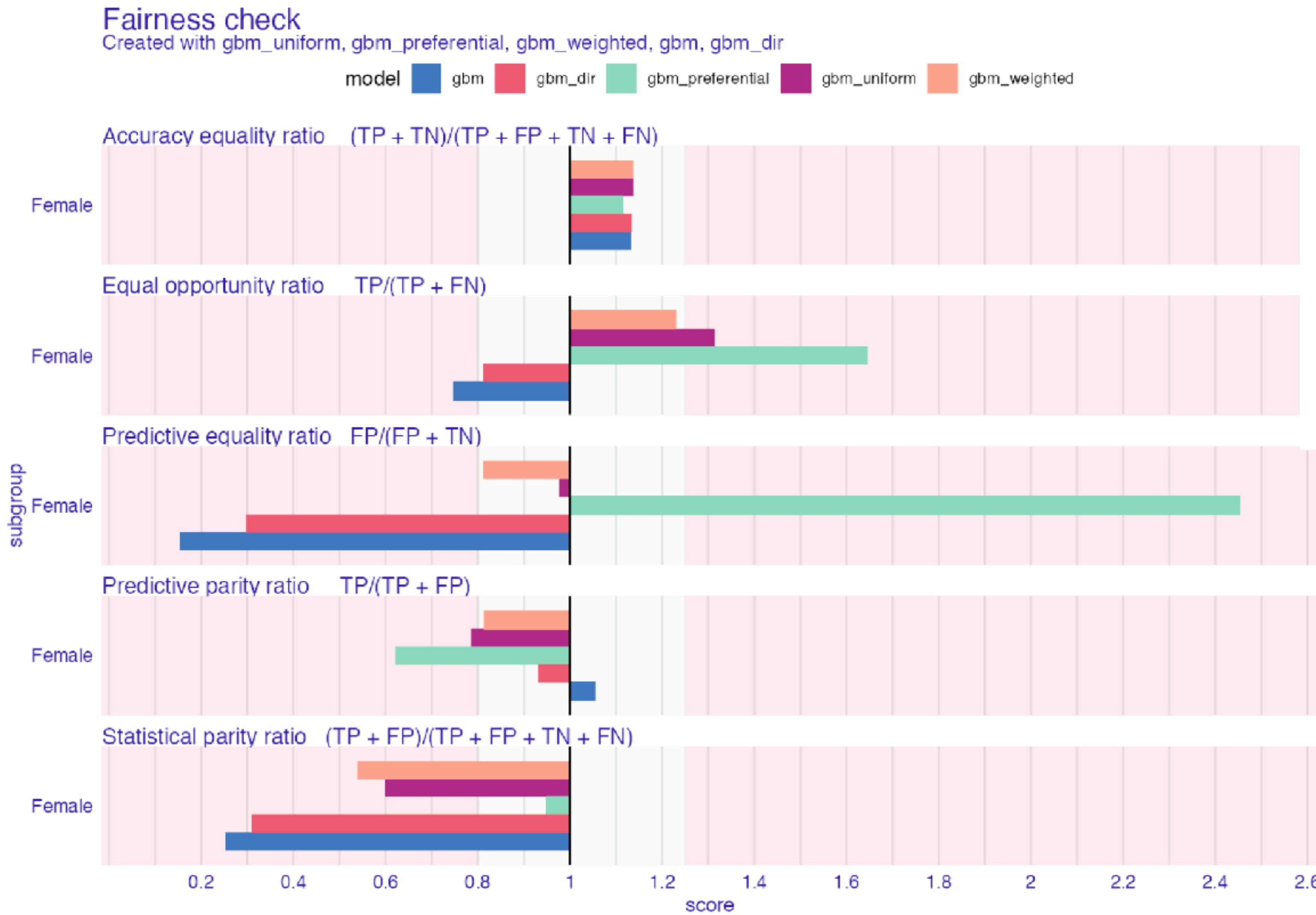
Created with ranger

model ranger



```
fobject <- fairness_check(fobject, gbm_explainer_u, gbm_explainer_p,
                           verbose = FALSE)

plot(fobject)
```



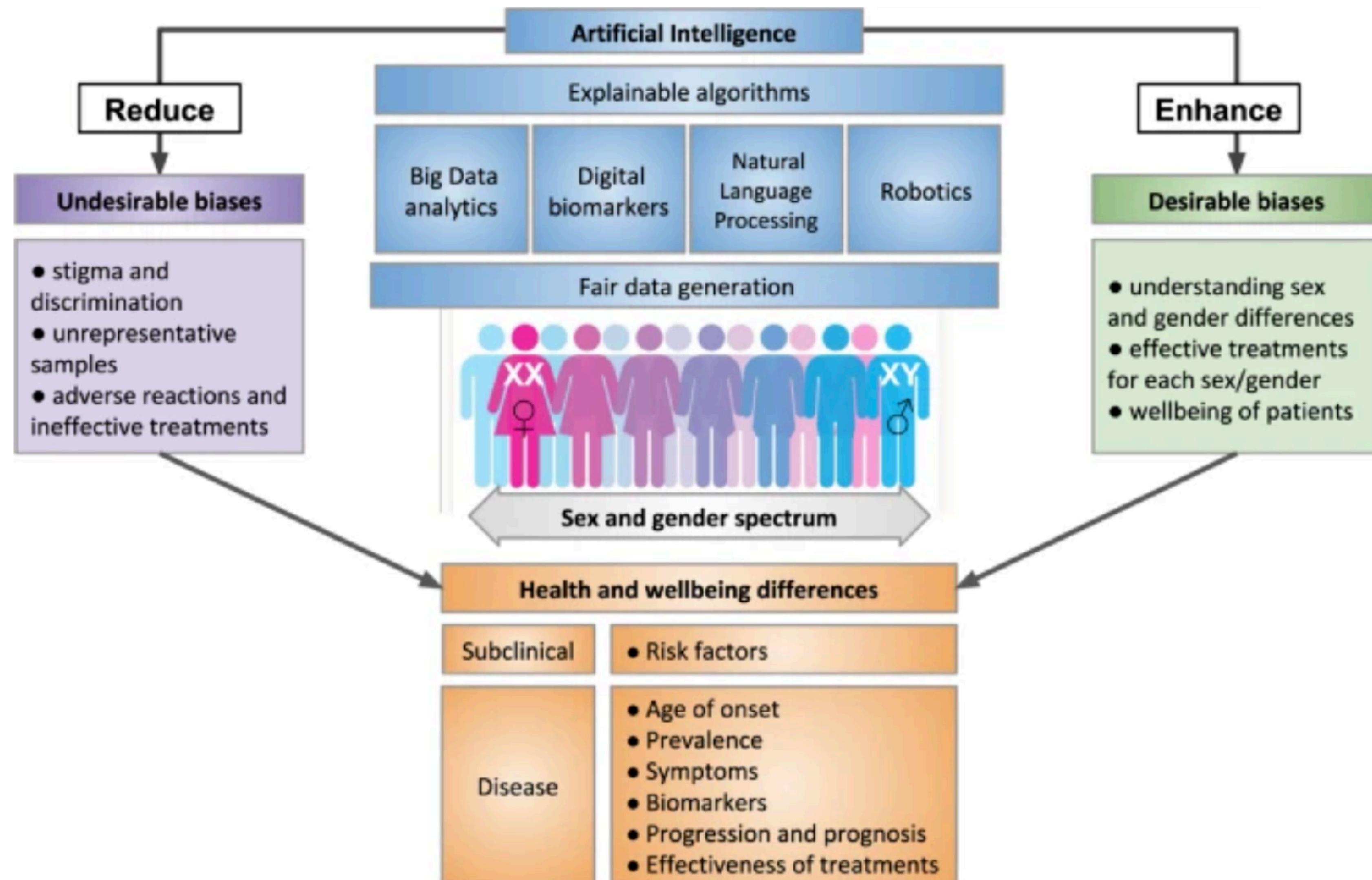
Model, explanation & bias

Bias mitigation strategies

Pre-processing techniques

Post-processing techniques

Fig. 2: Desirable and undesirable biases in artificial intelligence for health.



Transparency

Przejrzystość reguł

Model skriningowy Bacha do oceny ryzyka raka płuca

$$risk = \sum_{i=0}^n \left(1 - S_0^{exp(\beta X_i)} \right) \left(S_1^{exp(\beta X_i)} \right) \prod_{j < i} \left(S_0^{exp(\beta X_j)} \right) \left(S_1^{exp(\beta X_j)} \right)$$

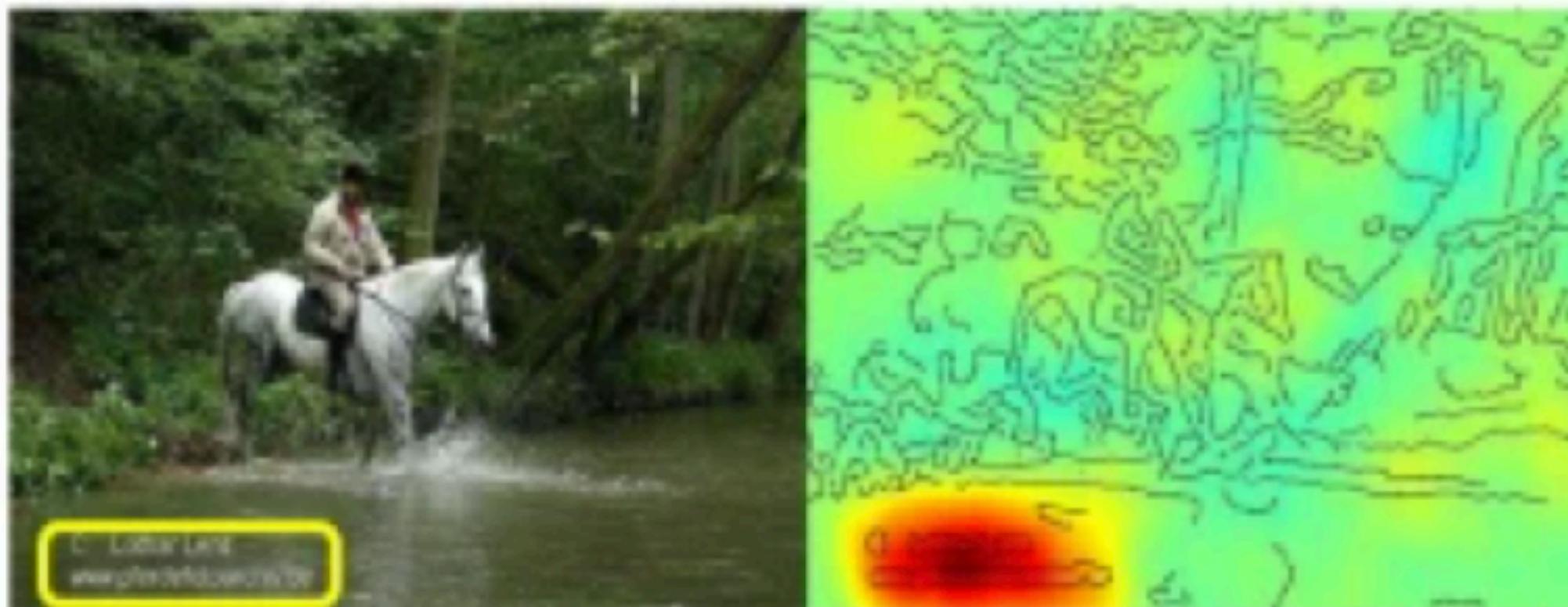
Bach PB, Kattan MW, Thornquist MD, Kris MG,
 Tate RC, Barnett MJ, et al. Variations in lung cancer risk
 among smokers. J Natl Cancer Inst. 2003;95:470–8.

Variable	Expression	Coefficient
<i>intercept</i>		-9.7960571
<i>age</i>		0.070322812
<i>age2</i>	$(age - 53.459001)^3 * I(age > 53)$	-0.00009382122
<i>age3</i>	$(age - 61.954825)^3 * I(age > 61)$	0.00018282661
<i>age4</i>	$(age - 70.910335)^3 * I(age > 70)$	-0.000089005389
<i>female</i>		-0.05827261
<i>qtyears</i>		-0.085684793
<i>qtyears2</i>	$(qtyears)^3$	0.0065499693
<i>qtyears3</i>	$(qtyears - 0.50513347)^3 * I(qtyears > 0)$	-0.0068305845
<i>qtyears4</i>	$(qtyears - 12.295688)^3 * I(qtyears > 12)$	0.00028061519
<i>smkyears</i>		0.11425297
<i>smkyears</i>	$(smkyears - 27.6577)^3 * I(smkyears > 27)$	-0.000080091477
<i>smkyears3</i>	$(smkyears - 40)^3 * I(smkyears > 40)$	0.00017069483
<i>smkyears4</i>	$(smkyears - 50.910335)^3 * I(smkyears > 50)$	-0.000090603358
<i>cpd</i>		0.060818386
<i>cpd2</i>	$(cpd - 15)^3 * I(cpd > 15)$	-0.00014652216
<i>cpd3</i>	$(cpd - 20.185718)^3 * I(cpd > 20)$	0.00018486938
<i>cpd4</i>	$(cpd - 40)^3 * I(cpd > 40)$	-0.000038347226
<i>asbestos</i>		0.2153936

Unmasking Clever Hans predictors and assessing what machines really learn

a

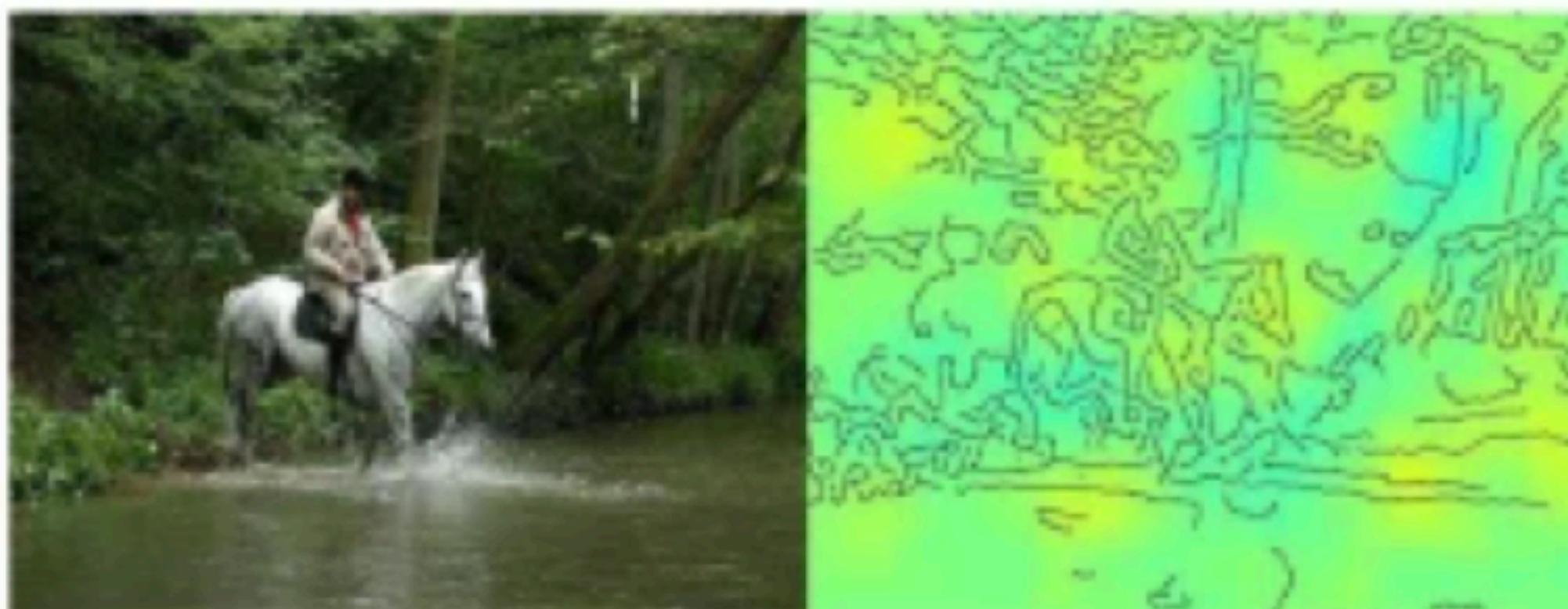
Horse-picture from Pascal VOC data set



Source tag
present



Classified
as horse



No source
tag present

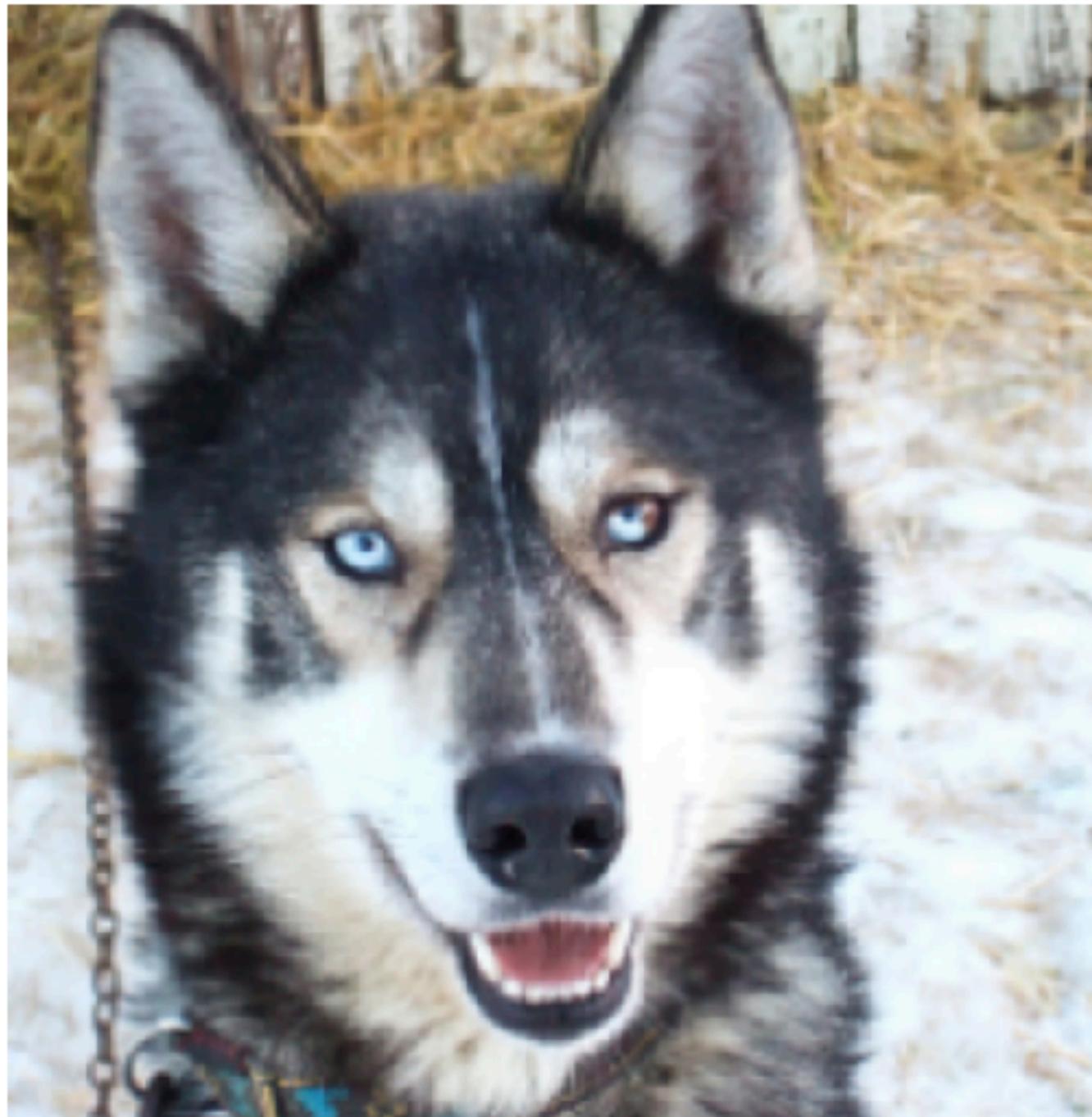


Not classified
as horse

Artificial picture of a car



“Why Should I Trust You?” Explaining the Predictions of Any Classifier



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

DeCoviD

Detection of Covid-19 related markers of pulmonary changes using Deep Neural Networks models
eXplainable Artificial Intelligence and Cognitive Compressed Sensing

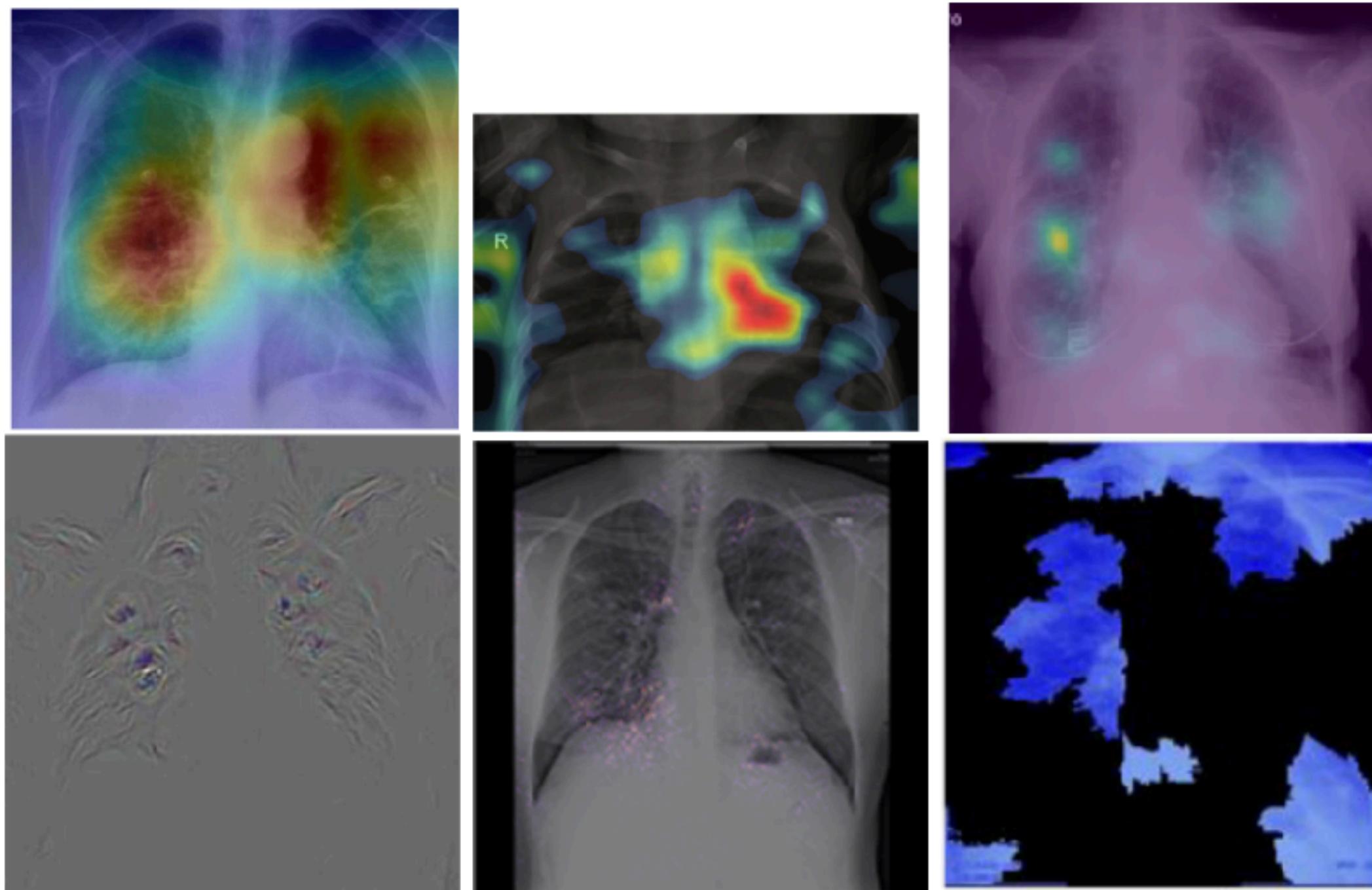


Fig. 4. Examples of XAI visualizations from studies: [1], [3], [27],

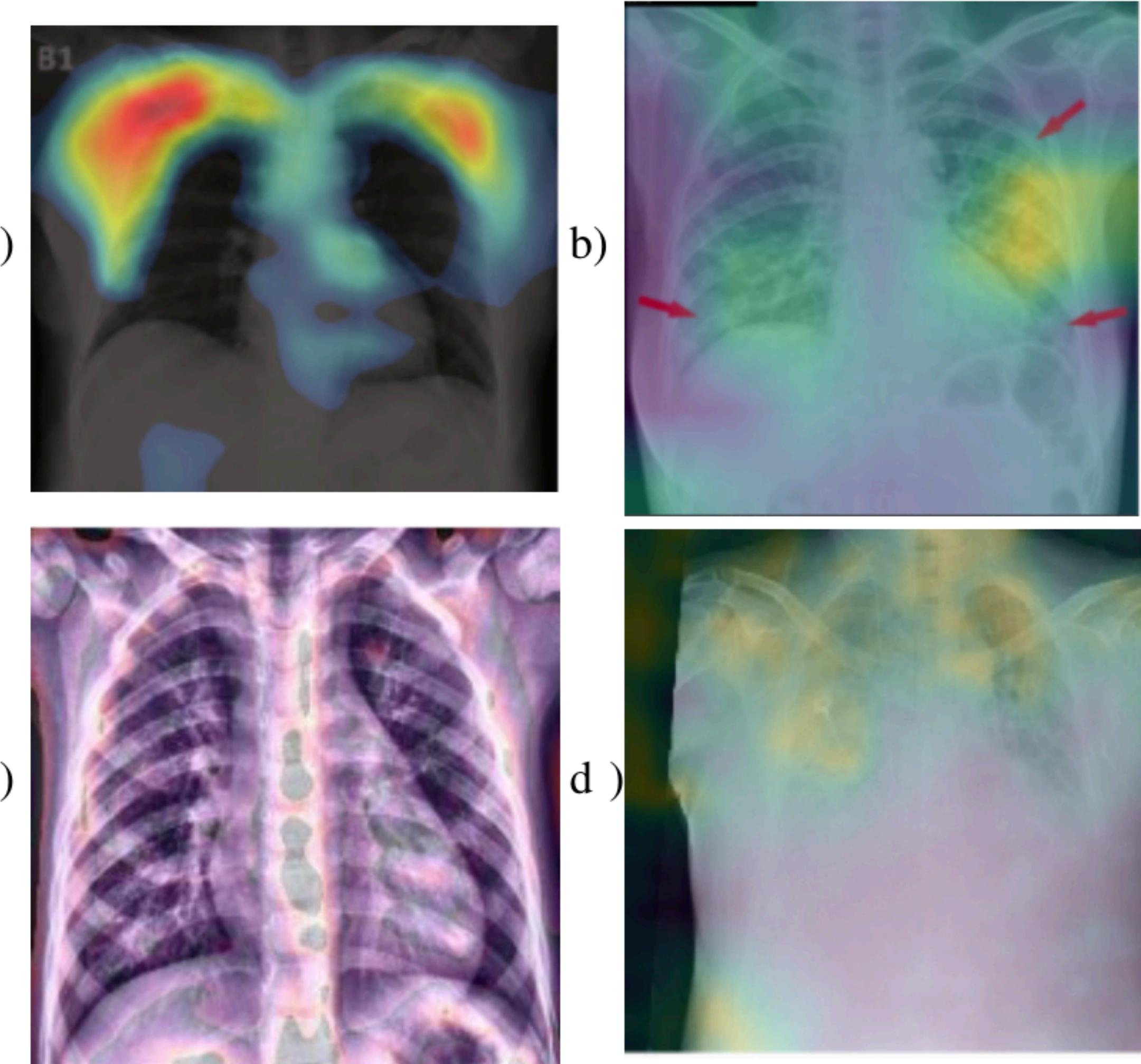
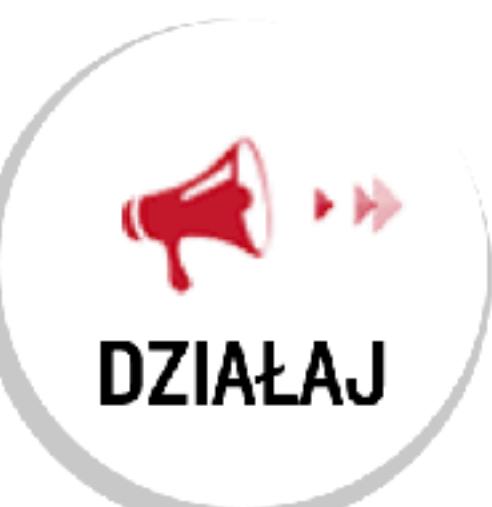


Fig. 5. Examples of biased model explanations [2], [13], [31], [66] Red arrows in the second image are marked by a radiologist to help localize the lesions. They were not present during model training.

INFORMACJE DZIAŁANIA WIEDZA →

Polecamy

SIN vs Facebook



Artykuł

Prawo do wyjaśnienia w pytaniach i odpowiedziach

05.04.2019

W kwietniu wejdzie w życie nowe prawo: do wyjaśnienia decyzji kredytowej. Na jaki problem odpowiada? Co konkretnie zmienia się w W jakich sytuacjach będzie można skorzystać z nowego prawa? Ile będzie to kosztowało? Odpowiadamy na najważniejsze pytania.

- 1. Na jaki problem odpowiada nowe prawo?**
- 2. Co konkretnie zmienia się w przepisach?**
- 3. W jakich sytuacjach mogę skorzystać z prawa do wyjaśnienia?**
- 4. Skąd mam wiedzieć, czy decyzja w mojej sprawie była automatyczna, czy podjęta przez człowieka?**
- 5. W którym momencie mogę skorzystać z prawa do wyjaśnienia?**
- 6. Zamierzam wziąć kredyt. Jak nowe prawo zmieni moją sytuację?**
- 7. Czy nowe prawo ułatwi mi uzyskanie kredytu?**
- 8. Co powinno znaleźć się w wyjaśnieniu udzielonym przez bank? Po czym poznam, że dał mi „dobrą” odpowiedź?**
- 9. Czy bank powinien ujawnić algorytm, na którym opiera się ocena mojej zdolności kredytowej?**
- 10. Ile zapłacę za skorzystanie z prawa do wyjaśnienia?**
- 11. Wyjaśnienie decyzji udzielone przez bank wydaje mi się niejasne, niepełne. Co mogę zrobić?**
- 12. Niepokoi mnie informacja otrzymana od banku (np. wydaje mi się, że bank popełnił błąd albo źle mnie ocenił). Co mogę zrobić?**
- 13. Bank nie chce udzielić mi wyjaśnienia. Co mogę zrobić?**
- 14. Jestem przedsiębiorcą. Co się zmieni w mojej sytuacji?**
- 15. Prowadzę działalność gospodarczą, ale staram się o kredyt w celach prywatnych. Czy mogę skorzystać z prawa do wyjaśnienia?**
- 16. Kiedy nowe prawo wejdzie w życie?**
- 17. Czy podobne przepisy obowiązują w innych krajach Unii Europejskiej?**
- 18. Co w tej sprawie zrobiła Fundacja Panoptikon?**

I. Na jaki problem odpowiada nowe prawo?

Do tej pory klient ubiegający się o kredyt w banku nie miał możliwości sprawdzenia, jakie jego dane zostały wzięte pod uwagę przy ocenie zdolności kredytowej i co przy tej ocenie miało znaczenie. Na przykład: czy o niskiej ocenie zdolności kredytowej przesądziło to, że za dużo pożyczam? Czy to, że za dużo wydaje na życie? A może tylko to, że kiedyś nie zapłacił w terminie rachunku za telefon?

Banki nadal nie mówią klientom całej prawdy. Czy komunikat KNF to zmieni?

19.10.2020



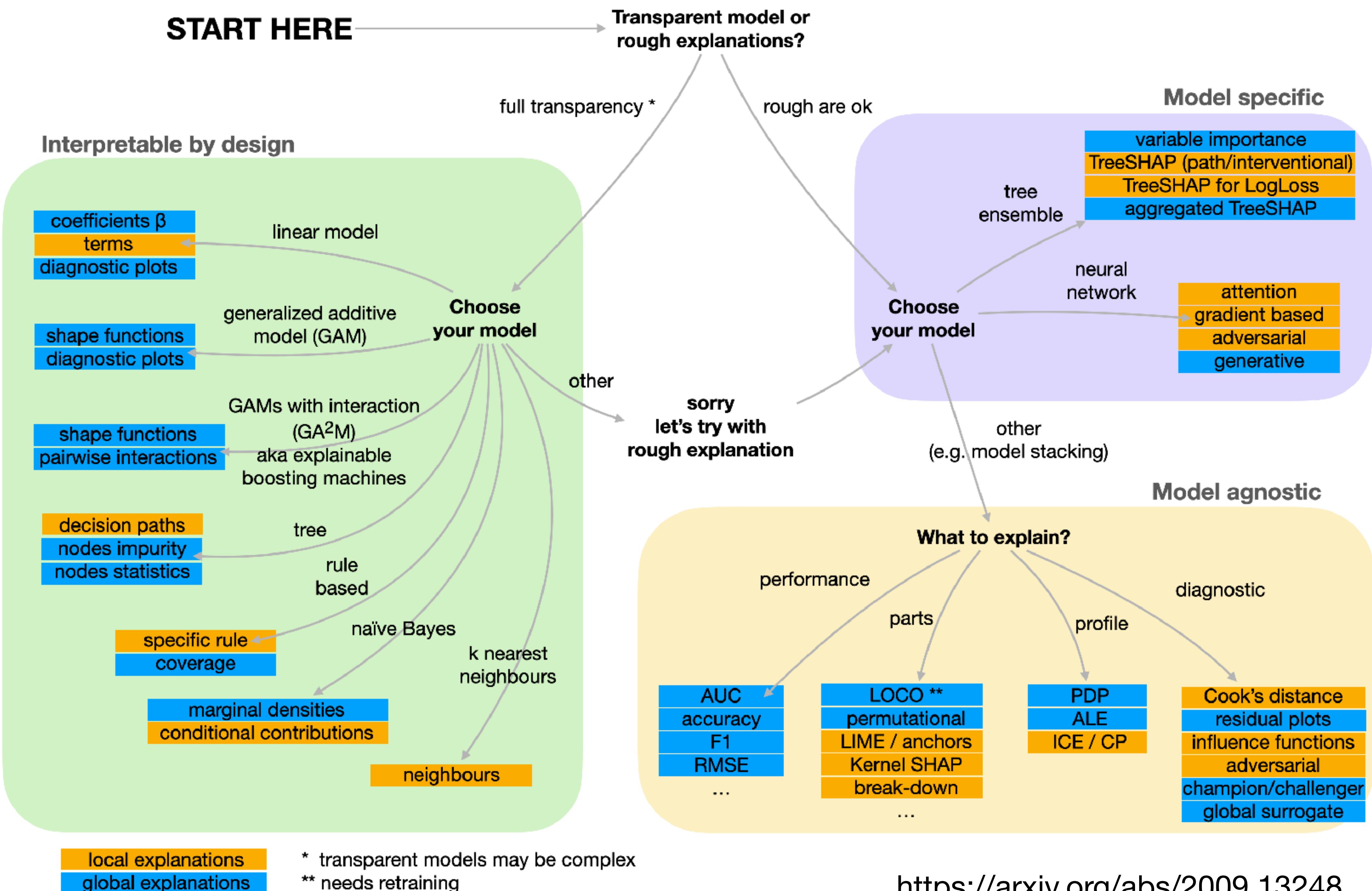
Kto obecnie stanowi większe ryzyko dla banku? Osoba pracująca w branży turystycznej czy przedsiębiorca prowadzący biznes gastronomiczny? Pandemia koronawirusa niejedno wywróciła do góry nogami, a zatem także porządek obowiązujący w bankach, które dużo ostrożniej podchodzą teraz do wniosków kredytowych. Ale czym konkretnie kierują się w swoich decyzjach? Możemy się domyślać, że z większą rezerwą traktują wnioski osób pracujących w branżach dotkniętych koronakryzysem. Takie domysły powinny jednak uciąć same banki, oferując wszystkim klientom rzetelne i pełne wyjaśnienie czynników, które wpłynęły na decyzję kredytową. W lipcu ów obowiązek, który w prawie bankowym pojawił się w związku z wdrożeniem RODO, [podkreśliła](#) Komisja Nadzoru Finansowego. Tego głosu w obronie prawa do wyjaśnienia banki nie mogą zignorować.

Mocne przepisy, słabe wykonanie

[Prawo do wyjaśnienia decyzji kredytowej](#) (por. art. 70a prawa bankowego) zostało wprowadzone w ramach dostosowywania polskich przepisów do RODO i obowiązuje od maja 2019 r. W założeniu miało zmniejszyć asymetrię informacyjną między pracownikami banków a osobami starającymi się o kredyt. Wcześniej ci drudzy mogli jedynie zgadywać, co konkretnie wpłynęło na ocenę ich zdolności kredytowej. Od maja ubiegłego roku mają jednak prawo uzyskać od banku szczegółową informację o wszystkich takich czynnikach. Ta informacja zwrotna jest bardzo cenna, choćby dlatego, że pozwala wychwycić i sprostować pomyłki, które przecież zdarzają się i w bankowych bazach danych, i w BIK-u. Dzięki niej klient lub klientka może też zrozumieć logikę, jaką posługuje się bank i – bez snucia domysłów – zaplanować swoje dalsze kroki. W czasach COVID-u mógłby lub mogłaby dowiedzieć się na przykład, jak bank patrzy na sektor, w którym pracuje, i czy czasem zmiana pracy nie jest najlepszą decyzją na drodze do pozyskania kredytu.

Niestety, z najświeższych przykładów, jakie do nas docierają, wynika, że banki nadal bronią się przed udzielaniem osobom starającym się o kredyt szczegółowej informacji zwrotnej. Z odpowiedzi, z którymi mieliśmy okazję się zapoznać, można jedynie wyczytać ogólne informacje: na przykład, że przy ocenie zdolności kredytowej bank wziął pod uwagę miejsce zatrudnienia i sektor gospodarki. Ale czy to był istotny czynnik wpływający na odmowę udzielenia kredytu? I czy zadziałał na korzyść, czy na niekorzyść klienta? Nie wiemy tego ani my, ani odbiorcy sztampowych „wyjaśnień”. To kolejny przykład pokazujący, jak przewrotnie banki interpretują nowe prawo do wyjaśnienia. Zamiast, jak wymaga tego prawo, wskazać konkretne czynniki – w tym dane osobowe – wpływające na ocenę zdolności kredytowej w określonej sprawie, serwują standardowe formułki, pełne nieprecyzyjnych i bardzo pojemnych sformułowanych.

START HERE



Kalkulator ryzyka Covid-19

Twoje ryzyko

Twoje ryzyko

Prawdopodobieństwa warunkowe są wyznaczone dla osoby zdiagnozowanej z chorobą covid-19 w wieku 30 lat i płci **męskiej**, **bez chorób towarzyszących**

W przypadku zachorowania warunkowe prawdopodobieństwo zgonu wynosi **0.1 %**
W przypadku zachorowania warunkowe prawdopodobieństwo hospitalizacji wynosi **13.1 %**

Wiek: 30

Płeć: Mężczyzna

Inne choroby: Nie

O modelu CRS-19

Skuteczność modelu

Model — Śmiertelność — Hospitalizacja

True positive rate

False positive rate

Co wpływa na wyliczoną warunkową śmiertelność?

Parametr	Wpływ na śmiertelność (%)
Średnia	+0.04
Wiek = 30	-0.03
Inne choroby = Nie	-0.005
Płeć = Mężczyzna	+0
Prognoza	+0.001

Jak warunkowa śmiertelność zależy od wieku?

Śmiertelność (%)

Wiek

Co wpływa na ryzyko hospitalizacji?

Parametr	Wpływ na ryzyko hospitalizacji (%)
Średnia	+0.29
Wiek = 30	-0.09
Inne choroby = Nie	-0.059
Płeć = Mężczyzna	-0.01
Prognoza	+0.13

Jak ryzyko hospitalizacji zależy od wieku?

Prawdopodobieństwo hospitalizacji (%)

Wiek

Accountability

Bezpieczeństwo i odpowiedzialność

Odpowiedzialne użycie danych

Podobnie jak energia jądrowa AI może być rozwijana w różnych modelach biznesowych.
Porównaj:

- Dodatkowe funkcjonalności urządzeń, ułatwiające operowanie urządzeniami (AI on device) takimi jak tomograf, ekspres do kawy czy telefon.
- Darmowe usługi w zamian za dane, które są przetwarzane i sprzedawane (Cambridge Analytica, systemy społecznościowe, „data is the new oil” itp).



Odpowiedzialne użycie danych

Handel niewolnikami

Przez wieki handel niewolnikami był traktowany jako coś zupełnie normalnego.

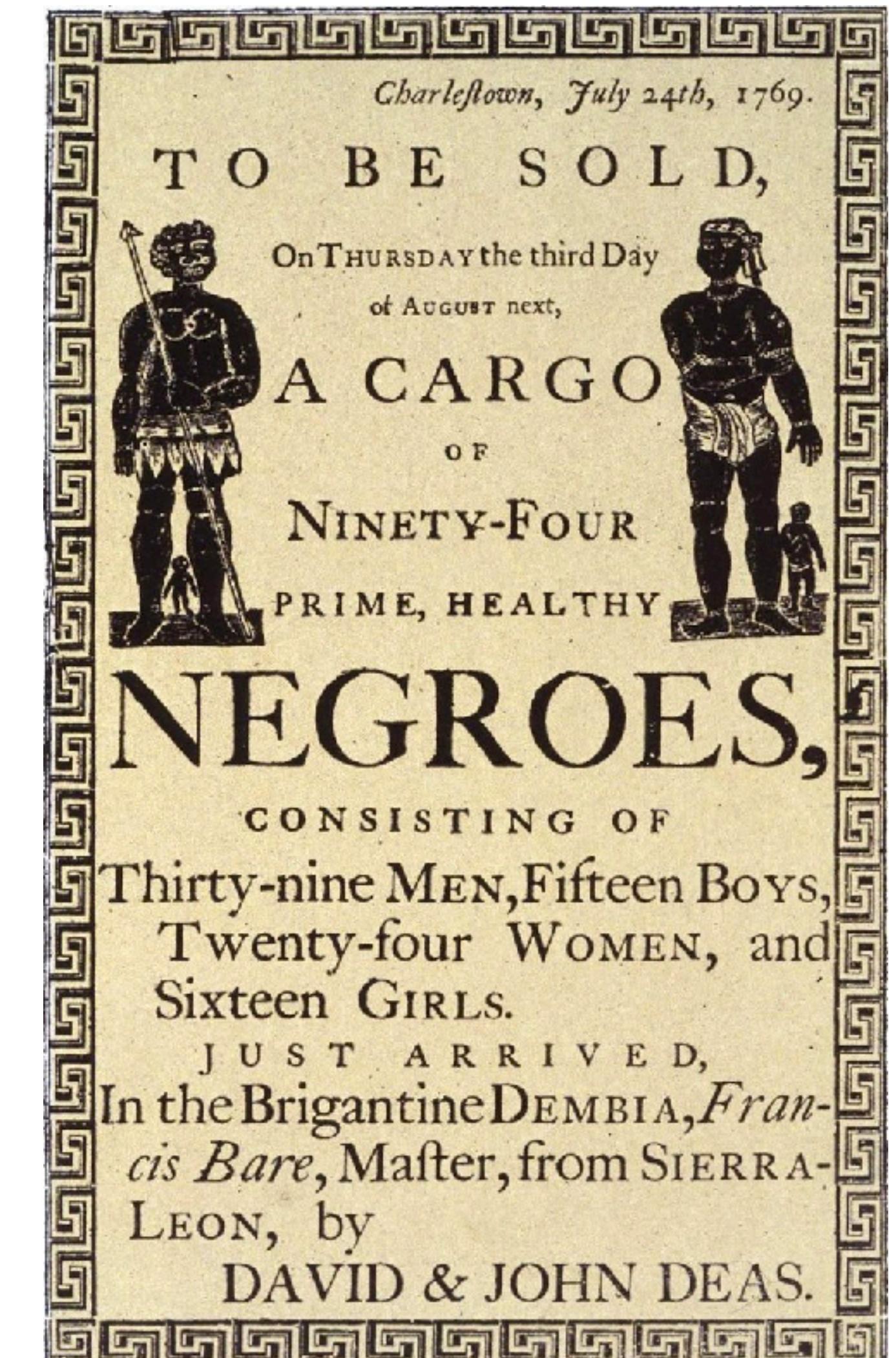
Niewolnicy byli własnością, którą można było sprzedać, wymienić, wycenić.

Dzisiaj niewolnictwo jest zakazane. Jest ono sprzeczne z prawami człowieka.

Handel danymi użytkowników

Nowe prawo Vermont wymaga, aby firmy, które kupują i sprzedają dane osobowe osób trzecich, były zarejestrowane. FastCompany stworzyło listę 121 brokerów danych działających w USA.

Przykłady: **wyszukiwanie ludzi:** Spokeo, ZoomInfo, White Pages, PeopleSmart, Intelius, PeopleFinders; **raportowanie kredytowe**, Equifax, Experian, TransUnion; **reklama i marketing**, Acxiom, Oracle, Innovis, KBM.



The Transformation of Work and the Law of Workplace Accidents, 1842-1910

John Fabian Witt

Moreover, in work accident cases courts put the nineteenth-century work ethic to dubious use by affirming the importance of worker control over their work and their working conditions in ways that placed the blame for accidents on the workers themselves—even in cases in which the injured employee could not have influenced or controlled the circumstances leading to the accident. By turning the work ethic into a moral imperative, the injured worker, like the unemployed worker, became morally suspect for failing to live up to his responsibility for the conditions of his work.⁵² In such cases, courts deflected the issue of employers' power over their employees by appealing to the work ethic and the notion that moral character inhered in the sound exercise of discretion in the workplace.⁵³

Odpowiedzialność za skutki wypadków



Michael Ramsey, a self-driving car expert with Gartner, characterized the video as showing "a complete failure of the system to recognize an obviously seen person who is visible for quite some distance in the frame. Uber has some serious explaining to do about why this person wasn't seen and why the system didn't engage."



MONITOR POLSKI

DZIENNIK URZĘDOWY RZECZYPOSPOLITEJ POLSKIEJ

Warszawa, dnia 12 stycznia 2021 r.

Poz. 23

**UCHWAŁA NR 196
RADY MINISTRÓW**

z dnia 28 grudnia 2020 r.

w sprawie ustanowienia „Polityki dla rozwoju sztucznej inteligencji w Polsce od roku 2020”

Rada Ministrów uchwała, co następuje:

§ 1. Ustanawia się „Politykę dla rozwoju sztucznej inteligencji w Polsce od roku 2020”, zwaną dalej „Polityką AI”, stanowiącą załącznik do uchwały.

§ 2. Wykonawcą Polityki AI jest minister właściwy do spraw informatyzacji.

§ 3. 1. Minister właściwy do spraw informatyzacji przedstawia Radzie Ministrów, w terminie do dnia 1 września danego roku, informację o realizacji działań w ramach Polityki AI za rok poprzedni.

2. Pierwszą informację o realizacji działań w ramach Polityki AI minister właściwy do spraw informatyzacji przedstawi Radzie Ministrów do dnia 1 września 2021 r.

§ 4. Uchwała wchodzi w życie z dniem następującym po dniu ogłoszenia.

<https://monitorpolski.gov.pl/M2021000002301.pdf>

Odpowiedzialność za szkody

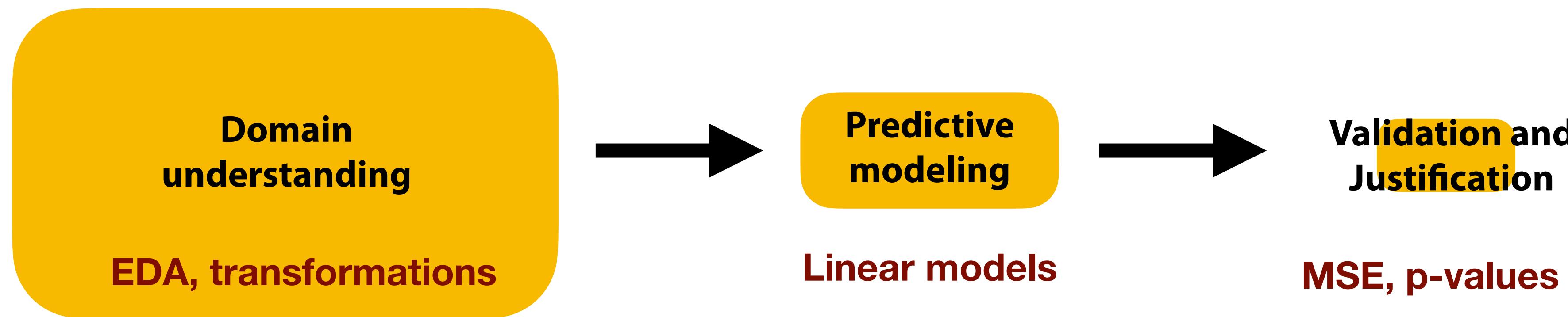
Tabela 2. Ramy polskiego ekosystemu AI

Ramy polskiego ekosystemu AI	
Wymiar	Kierunki działań politycznych
etyczny	<ul style="list-style-type: none">• godność ludzka i wsparcie autonomia człowieka wobec automatyki maszyn cyfrowych• globalny kodeks etyczny AI• godna zaufania AI
prawny	<ul style="list-style-type: none">• definicja legalna AI• przeciwdziałanie nadaniu osobowości prawnej AI• własność danych osobowych i ich przenaszalności• ochrona tajemnic przedsiębiorstwa i brak własności danych przemysłowych• własność intelektualna• odpowiedzialność za szkody wytwórców AI na zasadzie staranności, a operatorów AI na zasadzie ryzyka, a także rozróżnienie odpowiedzialności użytkowników końcowych od odpowiedzialności operatorów AI• wsparcie specyfikacji zamówień publicznych na rozwiązań AI oraz ułatwienie procesu zamawiania

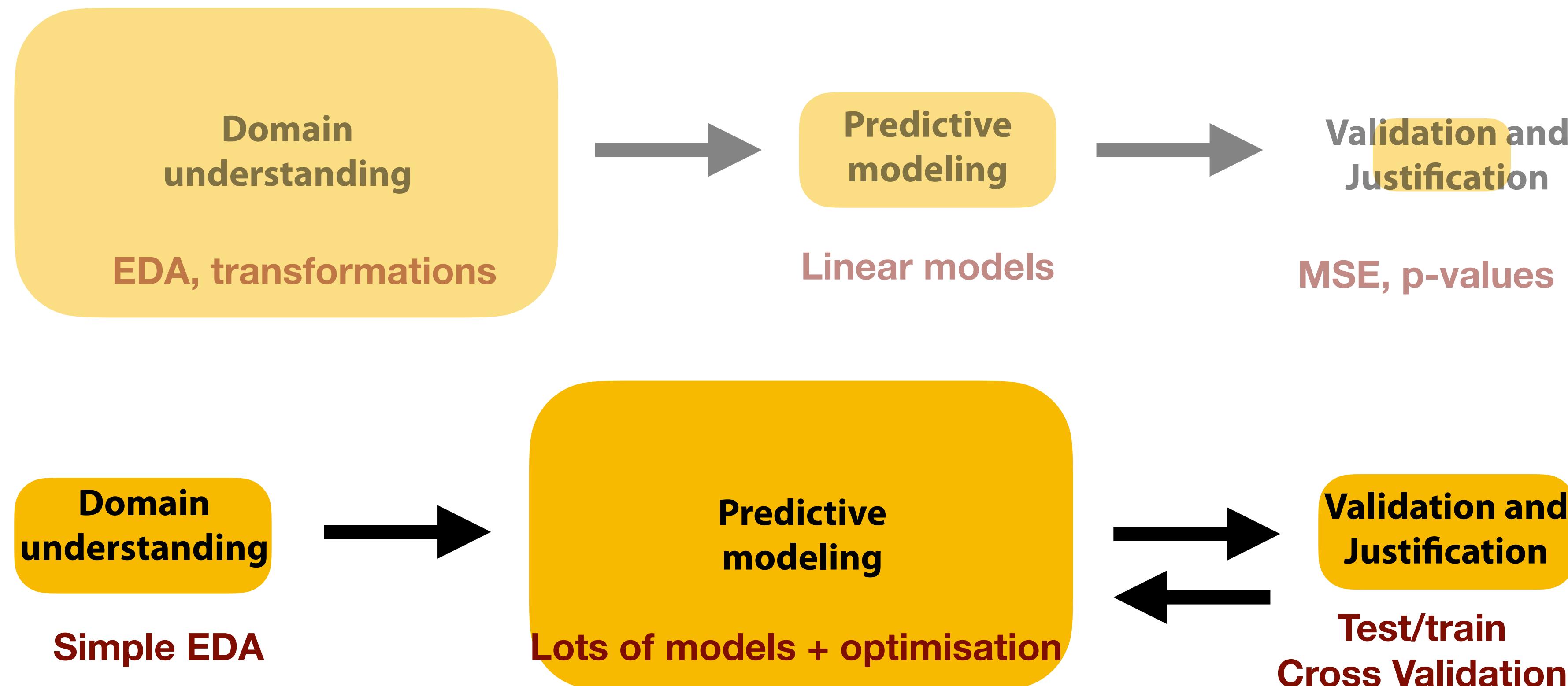
AI?

Jesteśmy na początku fascynującej podróży

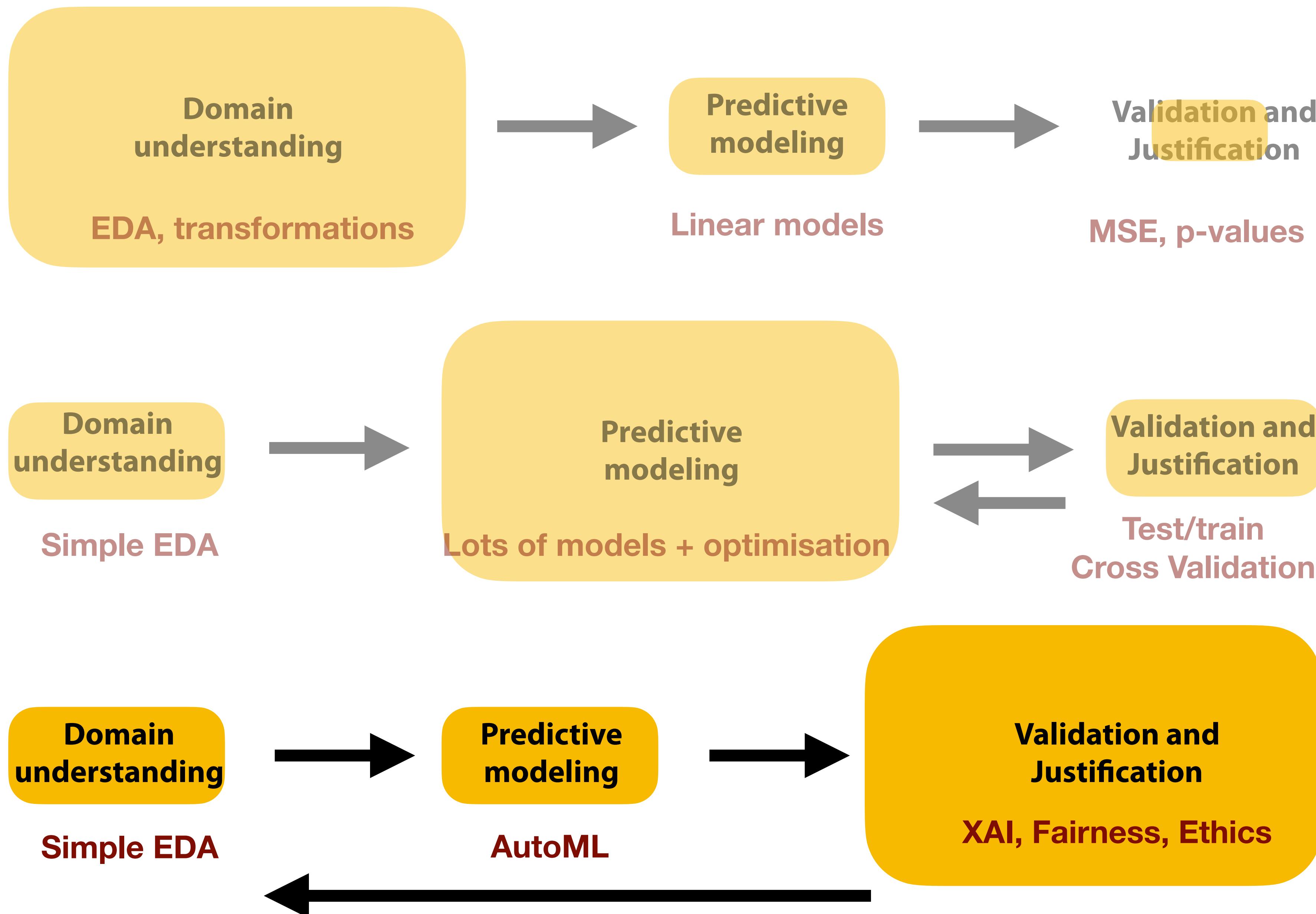
Statystyka - analiza oparta o głębokie zrozumienie dziedziny



Uczenie maszynowe - analiza oparta o dostępność danych i algorytmów

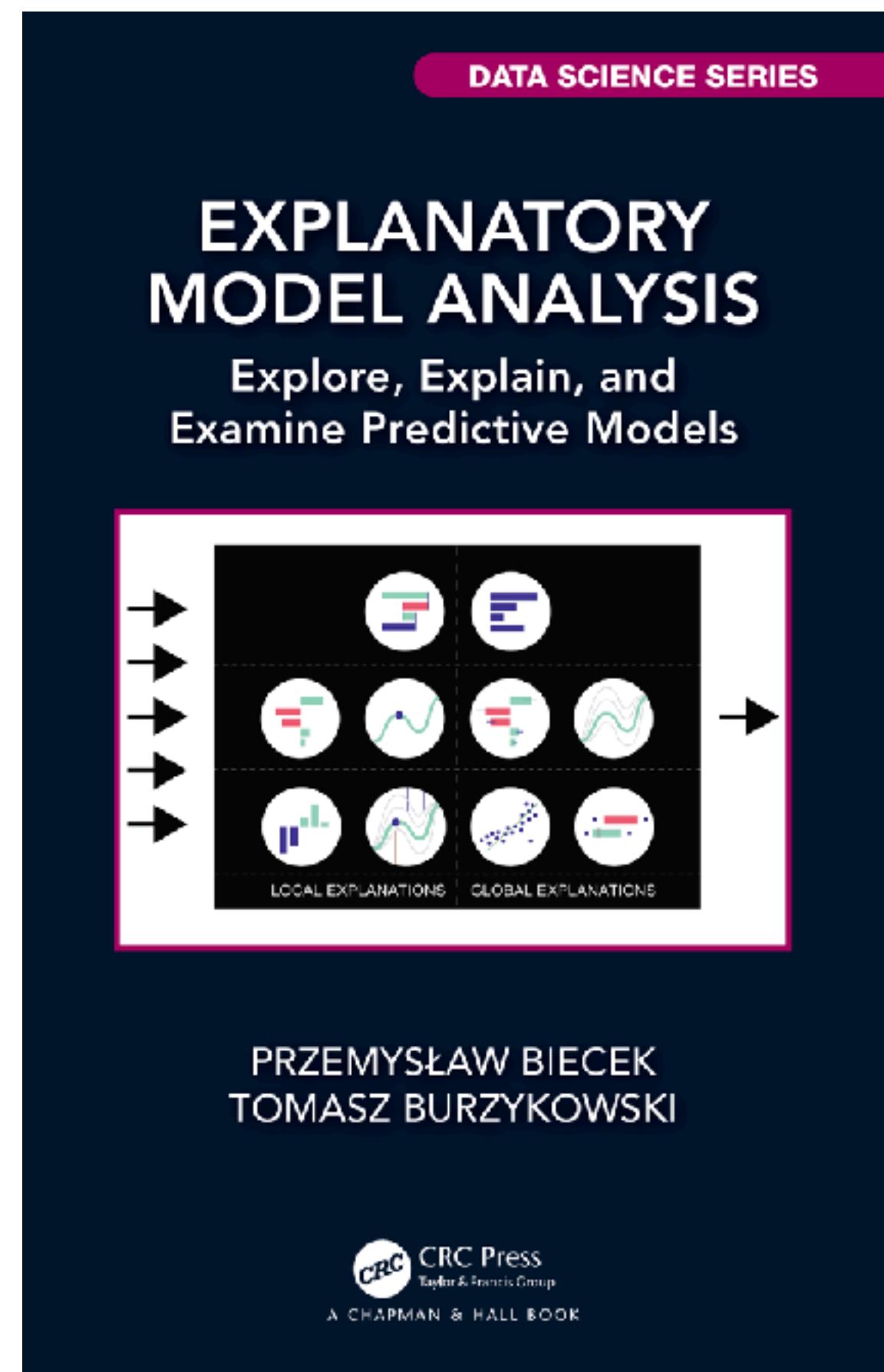


Modele oparte o dane uzgodnione z wiedzą dziedzinową

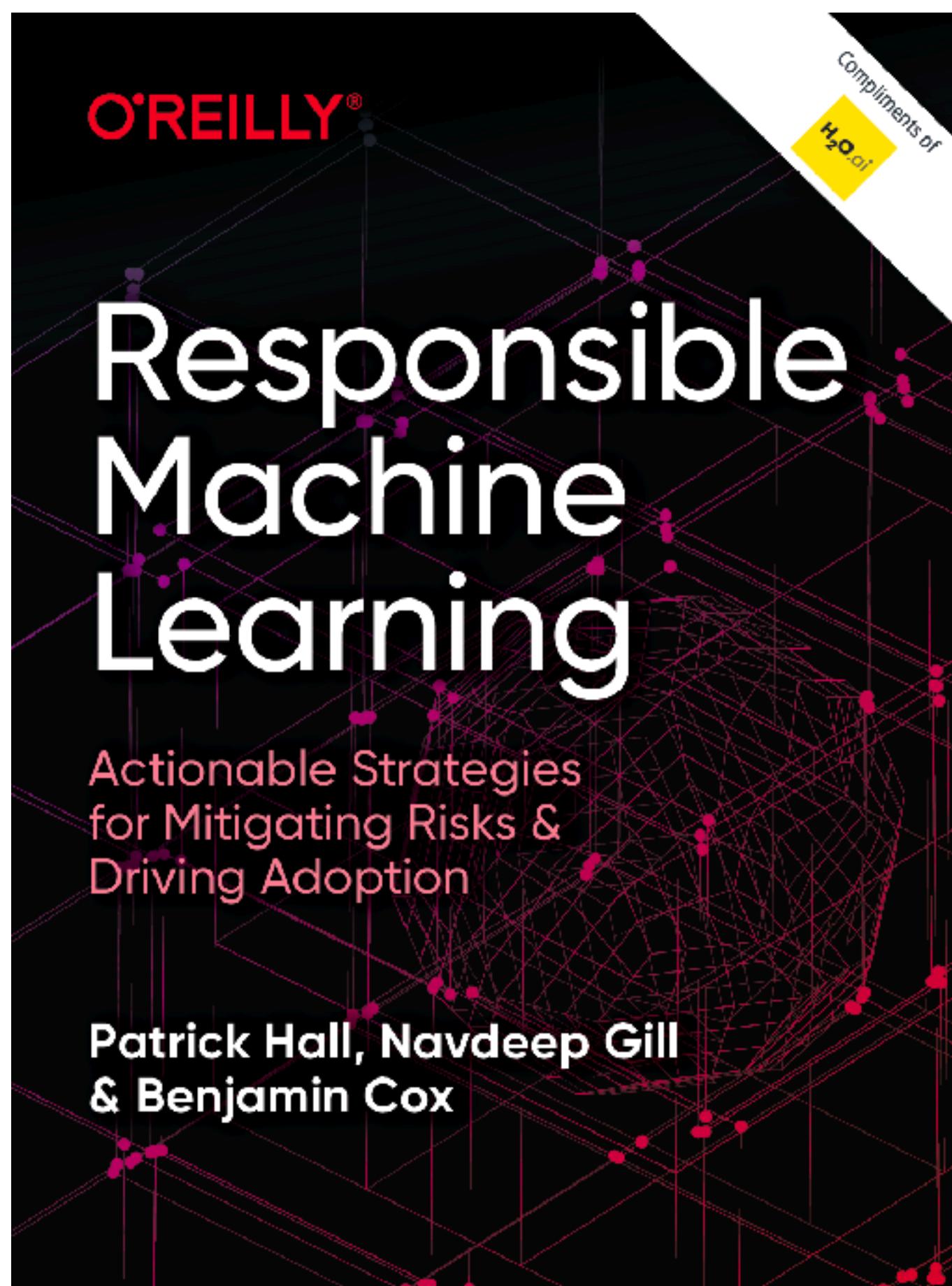


Literatura

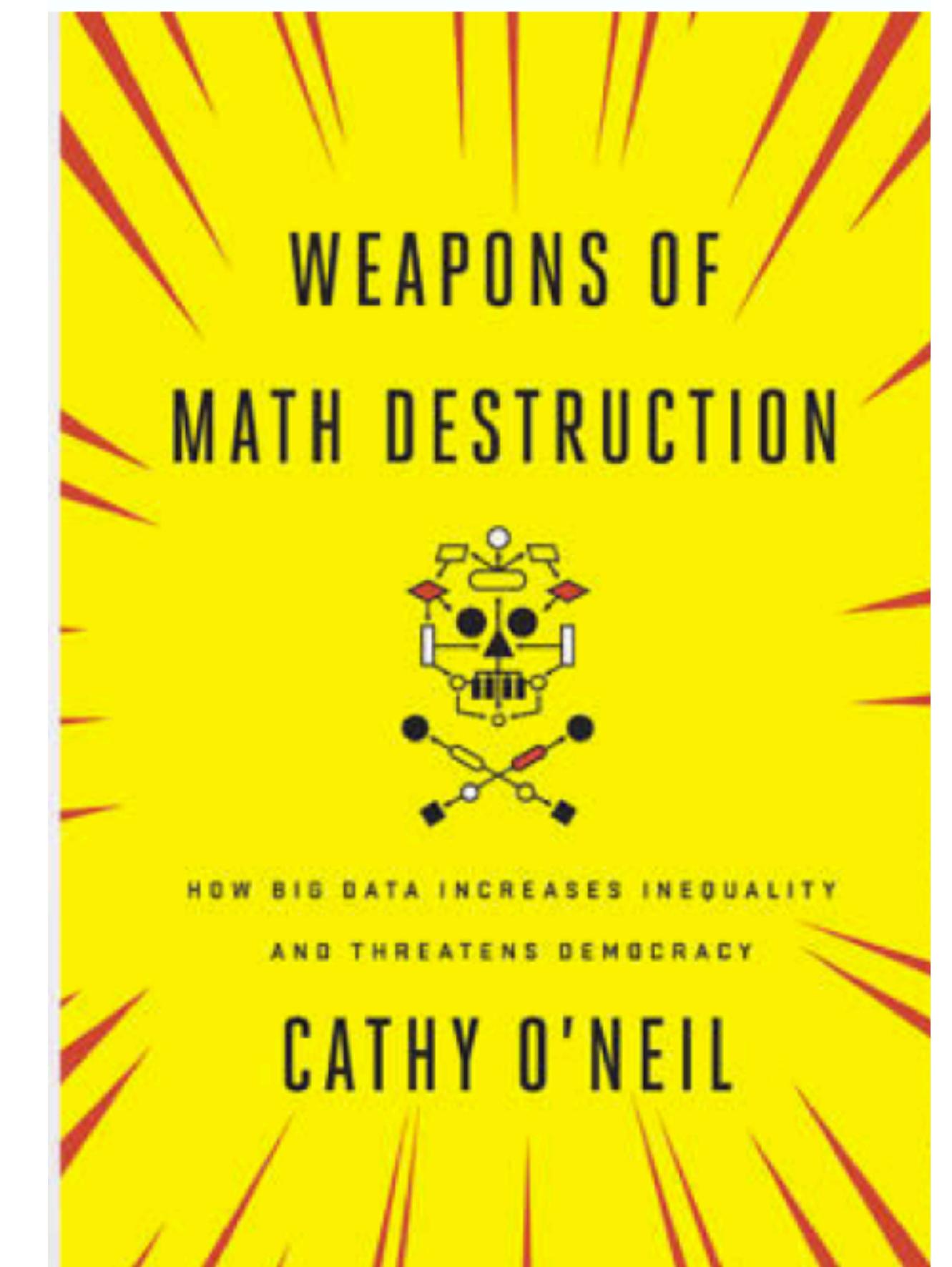
<http://ema.drwhy.ai/>



<https://www.h2o.ai/responsible-ai/>



https://en.wikipedia.org/wiki/Weapon_of_mass_destruction



<https://dalex.drwhy.ai/>

<https://medium.com/responsibleml>

*On a mission to responsibly build
machine learning predictive models*



Star 749

For R



R-CMD-check passing coverage 87%
CRAN 2.0.1 downloads 90K

Install from CRAN

```
install.packages("DALEX")
```

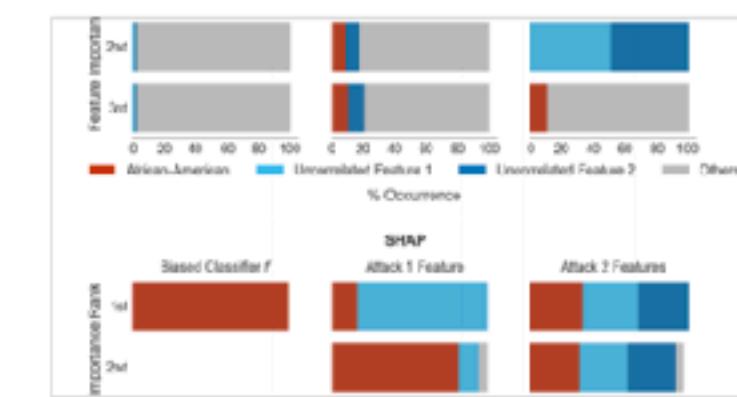
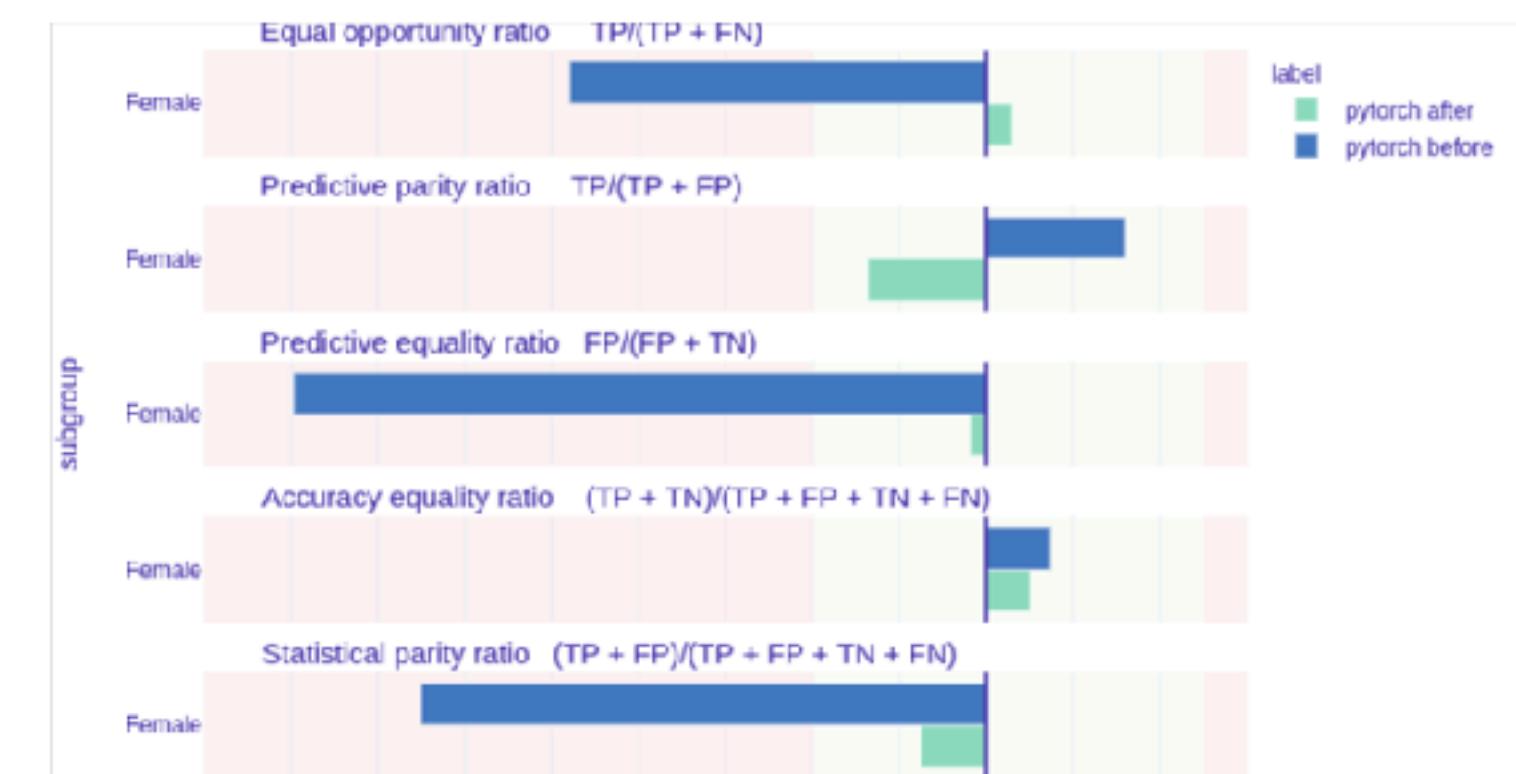
For Python



Python-check passing python 3.6 | 3.7 | 3.8
pypi package 1.0.0 downloads 16k

Install from PyPI

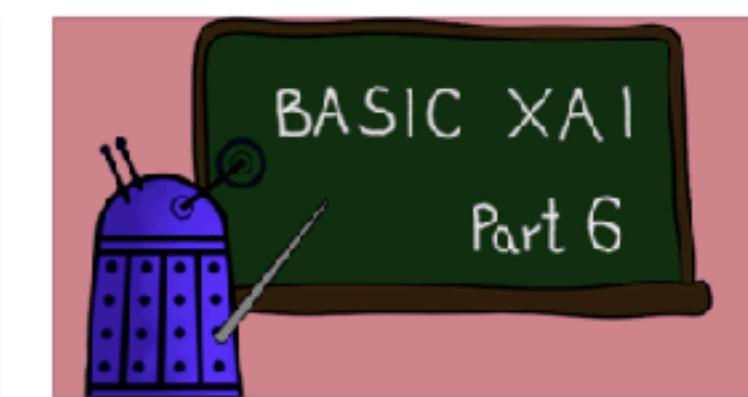
```
pip install dalex
```



**Be careful! Some models'
explanations can be fooled.**

Attack on LIME and SHAP explanations.

Wojtek Kretowicz
Jan 8 · 4 min read



**BASIC XAI with DALEX—Part
6: LIME method**

Introduction to model exploration with code examples for R and Python.

Anna Kozak
Jan 3 · 4 min read

**Visualize ML model bias
with dalex!**

Using python packages dalex and
fairtorch

Jakub Wiśniewski
Jan 12 · 5 min read



**Summarise the 2020 with R
and rgl**

The end of the year is a great time to summarize accomplishments of the team. This year in MI2DataLab we summarized good things that...

Przemysław Biecek
Dec 31, 2020 · 3 min read