MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

European University Institute | DEPARTMENT OF LAW

www.eui.eu

Справедливост при вземане на алгоритмични решения

Francesca Lagioia

Giovanni Sartor

European University Institute

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

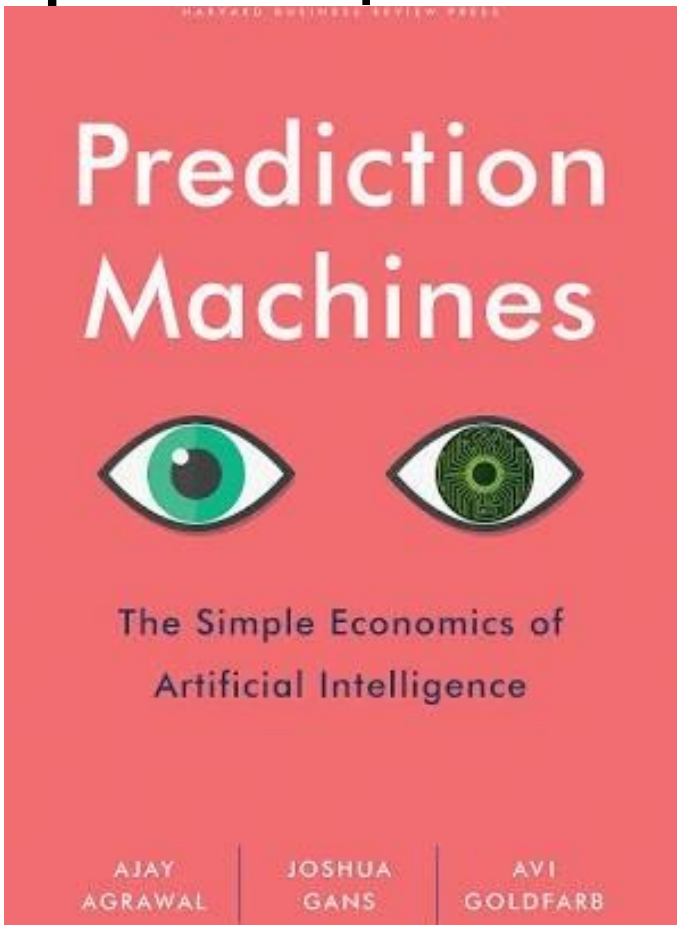
Съдържание

- ИИ при вземане на решения, засягащи индивиди
  - Възможни причини за несправедливост
- Принципът на справедливостта и неговото съдържателно измерение
- Несправедливост при ИИ
  - Предсказващата система COMPAS и случаят Лумис
  - Пример за играчка и критериите за оценка на справедливостта

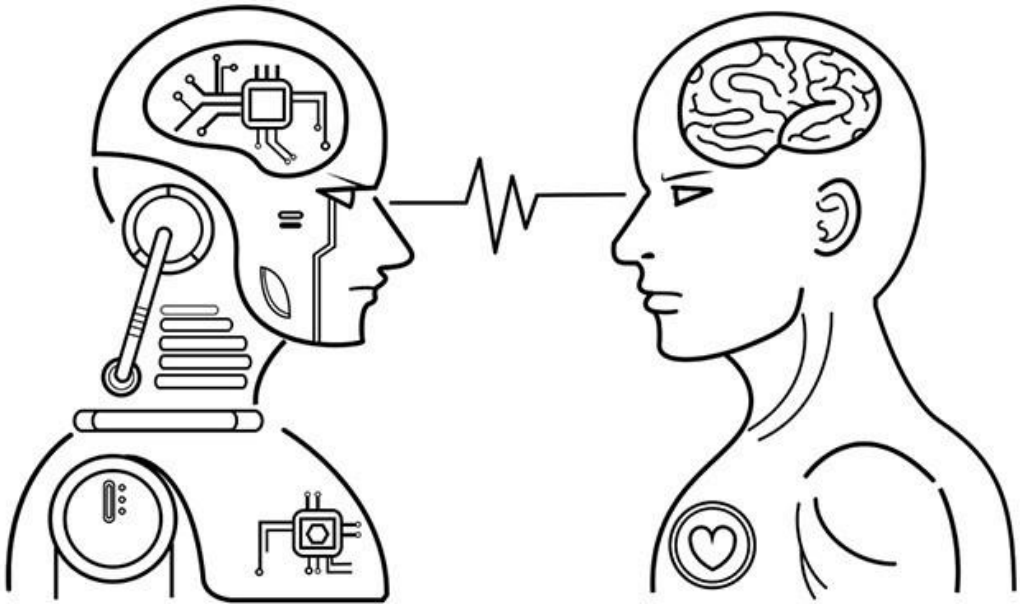
European University Institute
DEPARTMENT OF LAW

ИИ при вземане на решения относно индивиди: справедливост и дискриминация

- Комбинацията от AI и Big Data позволява автоматизирано вземане на решения дори в области, които изискват сложен избор, въз основа на множество фактори и непредефинирани критерии.
- През последните години се проведе широк дебат относно перспективите и рисковете от алгоритмични оценки и решения, засягащи индивиди



Дали системите с ИИ са по-добри от хората при оценяването ни?



В много области автоматизираните прогнози и решения са не само по-евтини, но и по-прецизни и безпристрастни от човешките.

- AI може да избегне типичните заблуди на човешката психология (прекомерна самоувереност, неприязън към загуба, закотвяне, пристрастия към потвърждението, евристика за представителност и т.н.) и широко разпространената човешка неспособност да обработва статистически данни, както и типичните човешки предразсъдъци (относно, напр., етническа принадлежност, пол, или социален произход).
- При много оценки и решения — относно инвестиции, набиране на персонал, кредитоспособност или също така и по съдебни въпроси, като освобождаване под гаранция, условно освобождаване и рецидив — алгоритмичните системи често се представят по-добре, според обичайните стандарти, от човешките експерти.

Или не?

Други подчертават възможността алгоритмичните решения да са погрешни или дискриминационни.

Само в редки случаи алгоритмите ще участват в изрична незаконна дискриминация, така нареченото различно третиране, основавайки своите резултати на забранени характеристики (предиктори) като раса, етническа принадлежност или пол.

По-често резултатът от дадена система ще бъде дискриминационен поради различното й въздействие, т.е. тъй като засяга непропорционално определени групи, без приемлива обосновка.

Системи, възпроизвеждащи силните и слабите страни на хората при вземането на преценки

Системите, базирани на контролирано обучение, могат да бъдат обучени на базата на минали човешки преценки и следователно могат да възпроизведат силните и слабите страни на хората, които са направили тези преценки, включително тяхната склонност към грешки и предразсъдъци.

- Например, система за набиране на персонал, обучена на минали решения за наемане, ще се научи да подражава на оценката на мениджърите за пригодността на кандидатите, вместо директно да прогнозира представянето на кандидата на работа. Ако минали решения са били повлияни от предразсъдъци, системата ще възпроизведе същата логика..

Предразсъдъци в набора за обучение

Предразсъдъците, вградени в наборите за обучение, могат да продължат да съществуват, дори ако входните данни (предикторите) към автоматизираните системи не включват забранени дискриминационни характеристики (напр. етническа принадлежност или пол).

Това може да се случи винаги, когато съществува корелация между дискриминиращи характеристики и някои предиктори

- Да предположим например, че предубеден мениджър по човешки ресурси не е нает кандидати от определен етнически произход и че хората с този произход живеят предимно в определени квартали. Обучителен набор от решения от този мениджър ще научи системите да не избират хора от тези квартали, което би довело до продължаване на отхвърлянето на заявления от дискриминираната етническа принадлежност. (Kleinberg et al (2019)).

Системи, предубедени срещу групи

В други случаи наборът за обучение може да бъде предубеден спрямо определена група, тъй като постигането на прогнозирания резултат (напр. представяне на работата) се приближава чрез заместител, който има различно въздействие върху тази група.

- Да приемем например, че бъдещото представяне на служителите (целта на интерес при наемане на работа) се измерва само с броя часове, отработени в офиса. Този критерий за резултат ще доведе до това, че предишното наемане на жени — които обикновено работят по-малко часове от мъжете, трябва да се справят със семейните тежести — се счита за по-малко успешно от наемането на мъже; въз основа на тази корелация (измерена на базата на предубеден пълномощник), системите ще предскажат по-лошо представяне на кандидат-жените.

Пристрастия на системата, вградени в предикторите

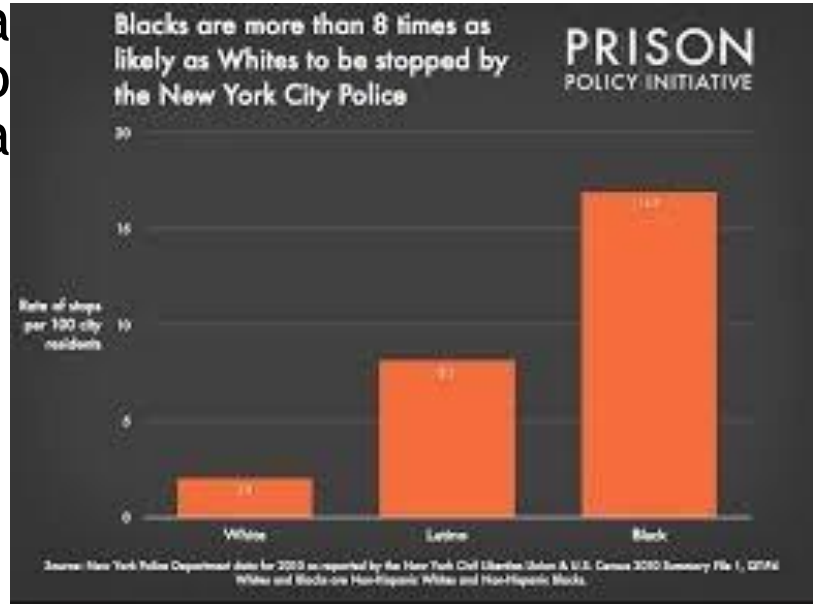
В други случаи грешките и дискриминацията могат да се отнасят до отклоненията на системата за машинно обучение, вградени в предикторите.

Една система може да работи несправедливо, тъй като използва благоприятен предиктор (характеристика за въвеждане), който се прилага само за членове на определена група (напр. фактът, че сте посещавали социално селективна институция за висше образование).

Несправедливостта може също да е резултат от приемането на пристрастни човешки преценки като предиктори (напр. препоръчителни писма).

Набор от данни, който НЕ отразява статистическия състав на съвкупността

И накрая, несправедливостта може да произтича от набор от данни, който отразява статистическия състав на населението.



- Да приемем например, че при молбите за освобождаване под гаранция или условно освобождаване, предишното криминално досие играе роля и че членовете на определени групи са обект на по-строг контрол, така че тяхната престъпна дейност да бъде по-често разкривана и да се предприемат действия. Това би означавало, че членовете на тази група като цяло ще получат по-неблагоприятна оценка от членовете на други групи, които се държат по същия начин.

- Членовете на определена група също могат да страдат от предразсъдъци, когато тази група е представена само от много малка част от набора за обучение,
- Това ще намали точността на прогнозите за тази група (например, разгледайте случая на фирма, която е назначавала няколко жени в миналото и която използва своите записи за минали наемания като свой набор за обучение).

Оспорване на несправедливостта на автоматизираното вземане на решения

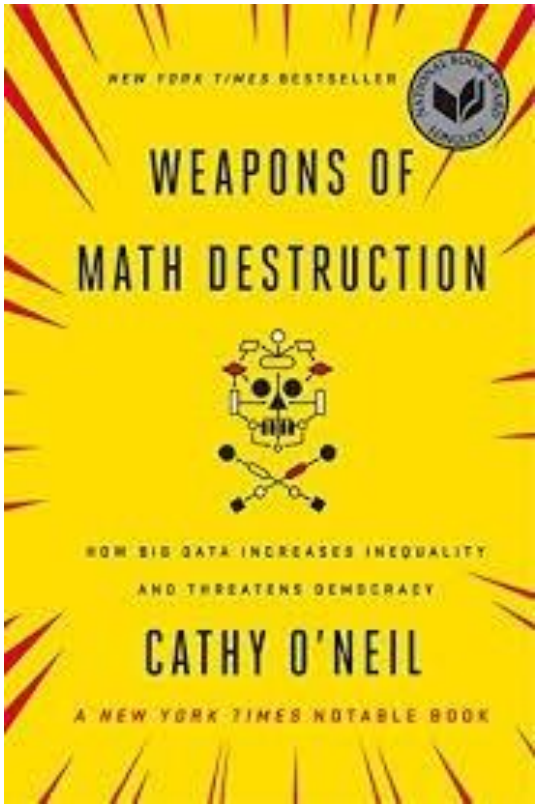
Беше отбелязано, че е трудно да се оспори несправедливостта на автоматизираното вземане на решения.

Предизвикателствата, повдигнати от засегнатите лица, дори когато са основателни, могат да бъдат пренебрегнати или отхвърлени, защото пречат на работата на системата, което води до допълнителни разходи и несигурност.

Всъщност прогнозите на системите за машинно обучение се основават на статистически корелации, срещу които може да е трудно да се спори въз основа на индивидуални обстоятелства, дори когато изключенията биха били оправдани.

Оръжия за унищожаване на математиката

“Алгоритъм обработва набор от статистически данни и извежда вероятност определен човек да е лошо нает, рисков кредитополучател, терорист или нещастен учител. Тази вероятност се дестилира в резултат, който може да преобърне нечий живот с главата надолу. И все пак, когато човекът отвърне на удара, „сугестивните“ изравнителни доказателства просто няма да го премахнат. Случаят трябва да е железен. Човешките жертви на оръжия за масово унищожение, както ще виждаме отново и отново, се държат на много по-висок стандарт на доказателства от самите алгоритми”. (O’Neil (2016))



Или не?

С подходящи изисквания, използването на алгоритми ще направи възможно по-лесното изследване и проучване на целия процес на вземане на решения, като по този начин ще направи много по-лесно да разберете дали е настъпила дискриминация. Чрез налагане на ново ниво на специфичност, използването на алгоритми също подчертава и прави прозрачни централни компромиси между конкуриращи се ценности. Алгоритмите са не само заплаха, която трябва да бъде регулирана; с правилните предпазни мерки, те имат потенциала да бъдат положителна сила за справедливост.

(Kleinberg, Ludwig, Mullainathan, e Sunstein (2018, 113)).

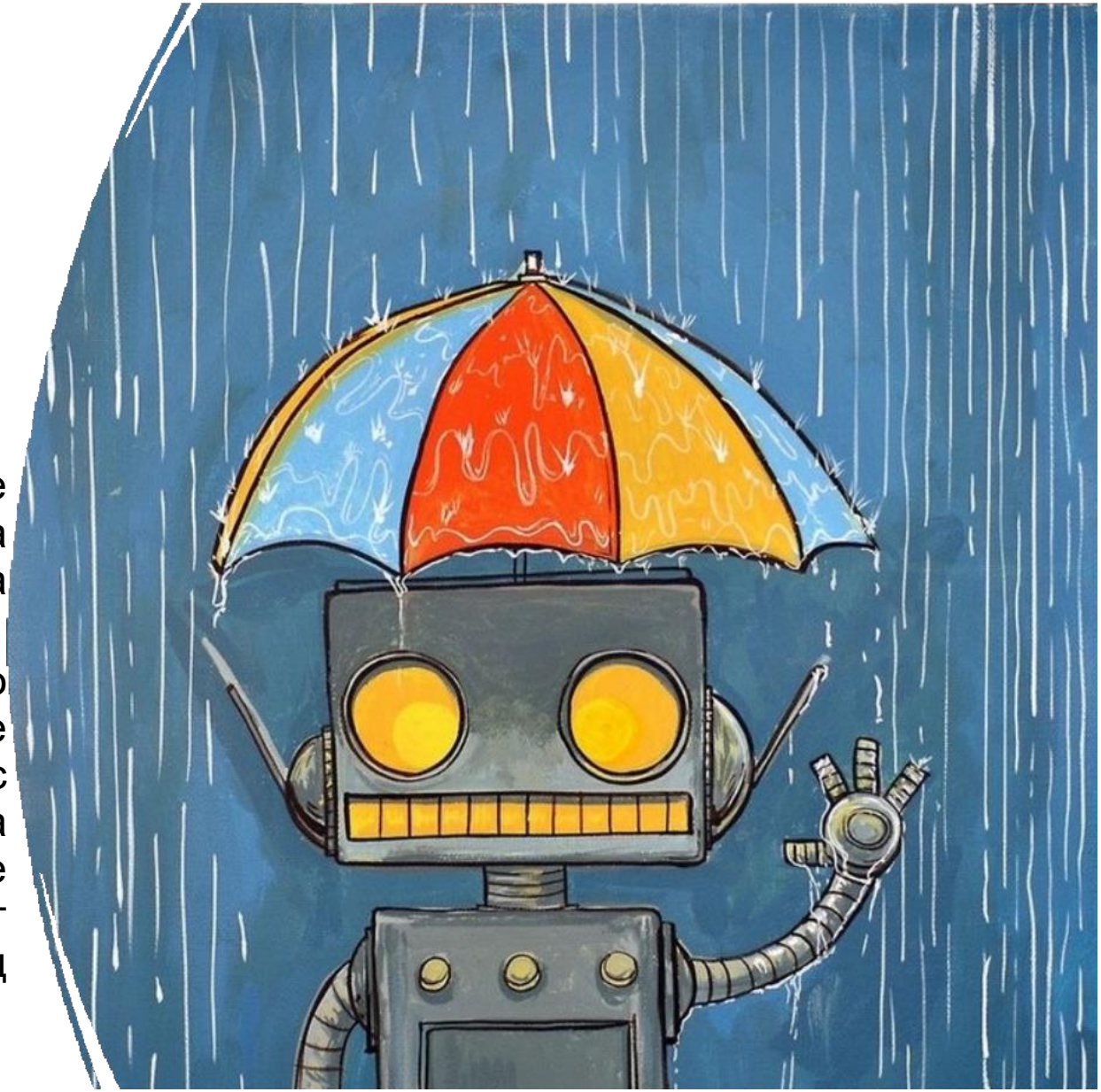
Оспорване на несправедливостта на автоматизираното вземане на решения

Тези критики бяха контрирани чрез наблюдението, че алгоритмичните системи, дори когато се основават на машинно обучение, са по-контролируеми от хората, вземащи решения, техните грешки могат да бъдат идентифицирани с точност и те могат да бъдат подобрени и проектирани, за да предотвратят несправедливи резултати.

Трябва ли да изключим използването на автоматизирано вземане на решения?

Изглежда, че току-що представените въпроси не трябва да ни карат да изключваме категорично използването на автоматизирано вземане на решения.

Алтернативата на автоматизираното вземане на решения не са перфектните решения, а човешките решения с всичките им недостатъци: една пристрастна алгоритмична система все още може да бъде по-справедлива от още по-пристрастния човек, вземащ решения.

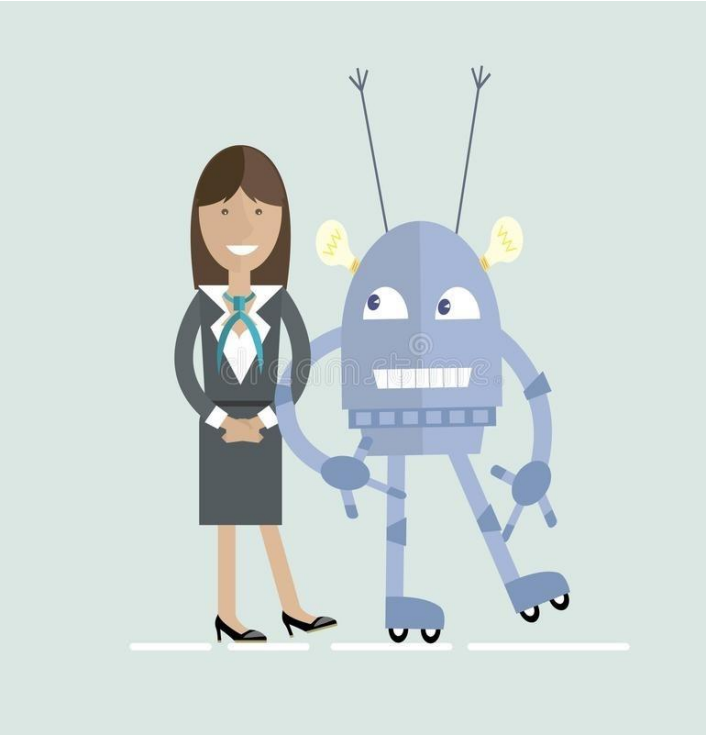


Хората + Алгоритми?

В много случаи най-доброто решение се състои в интегриране на човешки и автоматизирани преценки, като се даде възможност на засегнатите лица да поискат човешки преглед на автоматизирано решение, както и чрез предпочитане на прозрачността и разработване на методи и технологии, които позволяват на човешки експерти да анализират и преразгледат автоматизирано решение - изработка.

Всъщност системите с ИИ са демонстрирали способност да действат успешно и в области, традиционно поверени на обучената интуиция и анализ на хората, например медицинска диагноза, финансови инвестиции, отпускане на заеми и т.н.

Бъдещото предизвикателство ще се състои в намирането на най-добрата комбинация между човек и ИИ, като се вземат предвид капацитета и ограниченията и на двата.



Справедливост и ИИ

- Равно и справедливо разпределение на ползите и разходите
- Лица и групи, свободни от несправедливи пристрастия, дискриминация и стигматизация
- ИИ вземане на решения: информационна справедливост + справедливост на съдържанието на изводите/решението (избягване на предразсъдъци, дискриминация и т.н.)
  - подходящи математически или статистически процедури за профилиране,
  - технически и организационни мерки за гарантиране на коректността на личните данни
  - защитени лични данни (потенциални рискове, дискриминационни ефекти и др.)

Системата COMPAS: ИИ и несправедливост

- Актюерски инструмент за оценка на риска за определяне :
  - Риск от рецидив и подходящо коригиращо лечение
- Базирана на статистически алгоритми
- Нарушителите се класифицират в три категории: висок, среден и нисък риск
  - Тест с избираем отговор (137 въпроса)
  - Статични рискови променливи (напр. предишна криминална история, образование и др.)
  - Динамични рискови променливи (напр. злоупотреба с наркотици, трудов статус)

European University Institute
DEPARTMENT OF LAW

Случаят Лумис

- През 2013 г. Е. Лумис е обвинен в шофиране на откраднато превозно средство и бягство от полицията
- Окръжният съд нареди представяне на разследване, което включва оценка на риска COMPAS
- Лумис беше класифициран като висок риск от рецидив и осъден на 6 години затвор
- Решението беше обжалвано от Лумис за нарушаване на правата на текущ процес (напр. основни права на защита):
  - ❑ Функционирането на COMPAS не е известно
  - ❑ Валидността му не може да бъде проверена
  - ❑ Дискриминира по пол и раса
  - ❑ Базираните на статистика прогнози нарушават правото на индивидуално решение.

Случаят Лумис

През 2016 г. Върховният съд на Уисконсин отхвърли всички аргументи на ответника. Според Върховния съд :

- Статистическите алгоритми не нарушават правото на индивидуализирани решения
- Те трябва да се използват за „ подобряване на оценката на съдията на други доказателства при формулирането на индивидуална присъда
- Забрана решенията да се основават единствено на оценка на риска + задължение за мотивиране катозащита на правата на ответника.
- Отчитането на пола е необходимо за постигане на статистическа точност.
- Съдиите трябва да бъдат информирани за дебата относно расовата дискриминация на COMPAS

Предизвикателствата

През 2016 г. ProPublica публикува проучване (Larson et al. 2016):
Извадка: 11 757 обвиняеми, оценени от COMPAS (2013-2014 г.)
Цел: оценка на точността и коректността на COMPAS
Методология: Сравнение между прогнозирания процент на рецидивизъм и процента, който действително се е случил за период от 2 години.

Предизвикателствата

ProPublica резултати:

- Умерено-ниска точност на прогнозиране (61,2%)
- Предсказано е, че чернокожите обвиняеми са изложени на по-висок риск, отколкото в действителност. Вероятност за погрешна класификация с висок риск (45% чернокожи срещу 23% бели)
- Белите обвиняеми често се предвиждаха като по-малко рискови, отколкото бяха. Вероятност за погрешна класификация с нисък риск (48% бели срещу 28% чернокожи).

European University Institute
DEPARTMENT OF LAW

Опровержения

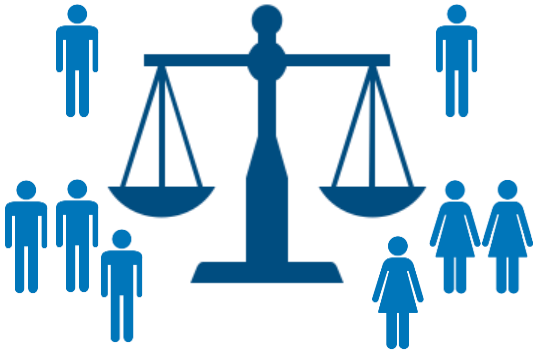
Според Northpoint (Dieterich et al 2016) ProPublica е допуснала няколко статистически и технически грешки

- Точността на прогнозите на COMPAS > точността на човешките преценки
- Общата скала за риск от рецидивизъм е еднакво точна за чернокожите и белите
- COMPAS спазва принципа на справедливостта
- Той не прилага расова дискриминация

Дебатът: Честен ли е COMPAS?

- Точени ли е?

- Честен ли е спрямо отделните хора?
- Честен ли е към групите?



European University Institute
DEPARTMENT OF LAW

Случаят SAPMOC

- 2000 обвиняеми
  - 1000 сини
  - 1000 залени
- Един единствен предсказател :
  - Ако има предишни престъпления, тогава вероятно ще рецидивира
- Предположение 1
  - предишни нарушители: 75% рецидивират
  - нарушители за първи път : 25% рецидивират
- Предположение 2
  - Сини: 75% предишни нарушители
  - Зелени: 25% предишни нарушители

SAPMOC Предположения

Реални резултати			
	рецидивизъм	без рецидивизъм	Общо
Предишни нарушители	750	250	1000
Без предишни нарушения	250	750	1000

SAPMOC предвиждания			
	рецидивизъм	без рецидивизъм	Общо
Предишни нарушители	1000	0	1000
Без предишни нарушения	0	1000	1000

	Позитивни	Негативни
	(TP+FN)/(TP+FN+FP+TN)	(TN+FP)/(TP+FN+FP+TN)
Сини	62.5%	37.5%
Зелени	37.5%	62.5%

	Позитивни	True Positives	False Positives	Негативни	True Negatives	False Negatives
	(TP+FP)	(TP)	(FP)	(TN+FN)	(TN)	(FN)
Сини	750	562.5	187.5	250	187.5	62.5
Зелени	250	187.5	62.5	750	562.5	187.5

SAPMOC точност

Точност	
(TP+TN)/(TP+FP+TN+FN)	
Сини	75,0%
Зелени	75,0%

SAPMOC справедливост

- Статистически паритет
- Равенство на възможностите
- Калибриране

- Грешка при условна употреба
- Равнопоставеност на отношението

Статистически паритет

- Всяка група трябва да има равен дял от положителни и отрицателни прогнози

Статистически паритет	Позитивни	Негативни
	(TP+FP)/(TP+FP+TN+FN)	(TN+FN)/(TP+FP+TN+FN)
Сини	75,00%	25,00%
Зелени	25,00%	75,00%

Равенство на възможностите

- Членовете на всяка група, които споделят едни и същи характеристики, трябва да бъдат третирани еднакво в равни пропорции.

Равенство на възможностите	Позитивни	Негативни
	TP/(TP+FN)	TN/(TN+FP)
Сини	90,0%	50,0%
Зелени	50,0%	90,0%

Калибриране

- Делът на правилните прогнози трябва да бъде равен във всяка група и по отношение на всеки клас.

Калибриране	Позитивни	Негативни
	TP/(TP+FP)	TN/(TN+FN)
Сини	75,0%	75,0%
Зелени	75,0%	75,0%

Грешка при условна употреба

- Съотношението между FP (FN) и общото количество положителни (отрицателни) прогнози трябва да бъде еднакво за двете групи.

Грешка	Positives	Negatives
	FP/(TP+FP) FN/(T	N+FN)
Сини	25,0%	25,0%
Зелени	25,0%	25,0%

Равнопоставеност на отношението

- Съотношението между грешките в положителните и отрицателните прогнози трябва да бъде еднакво във всички групи.

Равнопоставеност	Позитивни	Негативни
	FP/FN	FN/FP
Сини	300,0%	33,3%
Зелени	33,3%	300,0%

Сравнение SAPMOC/COMPAS?

- Еднаква точност в групите
- Различната основна ставка обяснява нарушаването на статистическия паритет, равенството в третирането и равните възможности
- Нарушаването на критериите за справедливост не води непременно до несправедливост
- Да наложим ли статистически паритет? (По-ниска точност + по-висок фалшив процент + дискриминация срещу индивиди)
- Индивидуална справедливост срещу групова справедливост

Справедливостта при автоматизирано вземане на решения

- Разопаковане на решението
  - Несправедливост в прогнозирането (забранени функции, набор от данни с предубедени данни, прокси сървър с предубедени данни и т.н.)
  - Несправедливо класифициране (праг – утвърдителни действия)
  - Несправедливост в решението (оптимизиране на правото/ценностите)

- Прогнозните системи като инструменти за разбиране на реалността

Поглед в бъдещето

- ИИ твърде често се възприема като източник на заплахи, а правото твърде често се възприема като трудно и понякога дори недостъпно за гражданите
- Комбинацията от ИИ и закон може да бъде ключът към защитата на гражданите и да направи закона достъпен за широката общественост

MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe
Co-financed by the European Union Connecting Europe Facility
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423