



Разбиране на алгоритмичната пристрастност

Идентифициране и намаляване на пристрастиата в данните и алгоритмите на изкуствения интелект с цел насърчаване на справедливостта и равнопоставеността





Цели на обучението

Накратко



- Да се разбере какво е алгоритмична пристрастност и как възниква.
- Да се идентифицират различните видове пристрастия в данните и системите за изкуствен интелект.
- Да се разпознават рисковете, свързани с използването на ИИ в образователния контекст.
- Да се анализират въздействието на пристрастието върху равнопоставеността и приобщаването.
- Прилагане на основни стратегии за намаляване на пристрастието в преподавателската си практика.



Ключови понятия и дефиниции

ОСНОВНИ КОНЦЕПЦИИ ЗА ИДЕНТИФИЦИРАНЕ И АДРЕСИРАНЕ НА ПРЕДРАЗСЪДЪЦИТЕ В СИСТЕМИТЕ ЗА ИЗКУСТВЕН ИНТЕЛЕКТ

Алгоритмична пристрастност

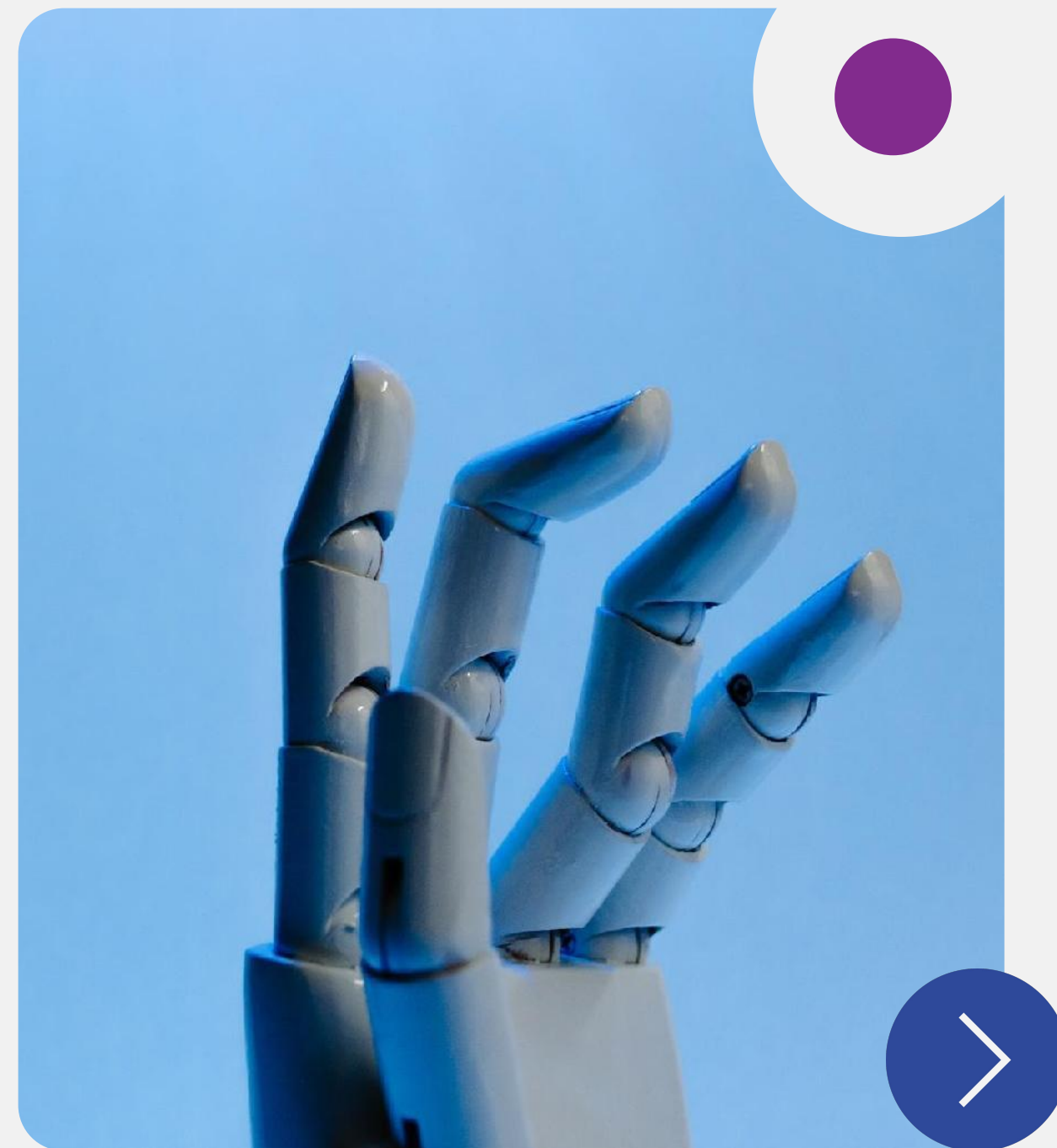
Алгоритмична пристрастност възниква, когато система за изкуствен интелект генерира систематично несправедливи резултати поради пристрастни данни, дизайнерски решения или имплицитни предположения

Представителна пристрастност

Представителна пристрастност възниква, когато определени групи са недостатъчно или прекомерно представени в обучителните данни.

Историческа пристрастност

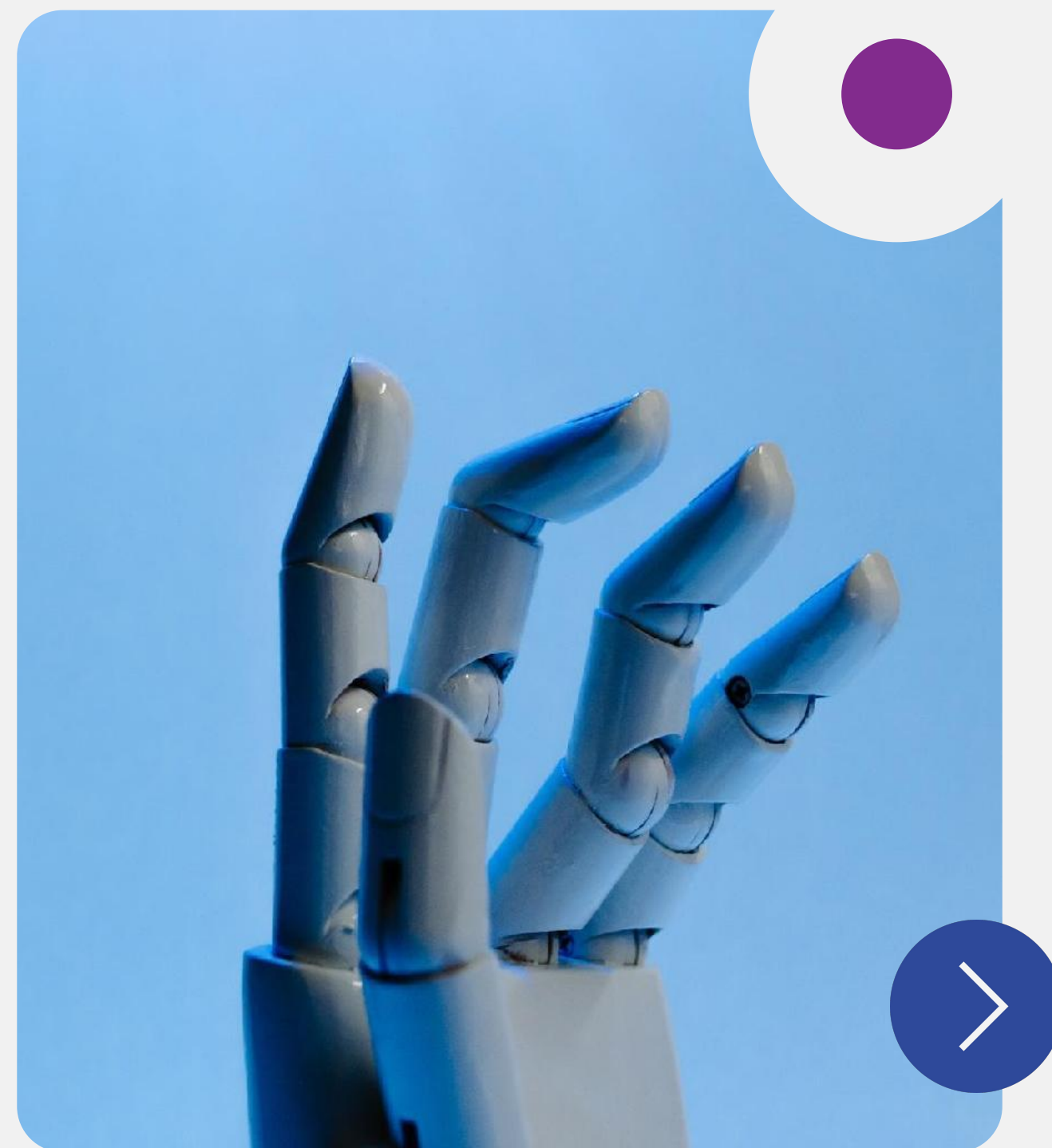
Историческият пристрастия възниква, когато наборите от данни отразяват минали неравенства, които се възпроизвеждат или усилват от системите за изкуствен интелект



Какво е алгоритмична пристрастност?

*Как възникват изкривявания в данните и
алгоритмите на изкуствения интелект*

- Алгоритмичната пристрастност възниква, когато дадена система за изкуствен интелект генерира систематично несправедливи или неравностойни резултати поради данните, с които е обучена, направените дизайнерски избори или начина, по който интерпретира информацията.
- То не винаги е умишлено, но може да възпроизведе или усили съществуващите неравенства.
- Алгоритмичната пристрастност не е само технически проблем, а и отражение на по-широки социални неравенства, заложиени в данните и институционалните практики.



Основни видове пристрастност



Исторически пристрастия



Когато данните за обучение отразяват минали неравенства, които се вграждат в системата за изкуствен интелект.

Пристрастност при агрегиране



Когато един модел се прилага към различни групи, без да се отчитат разликите.

Пристрастност при представянето



Когато определени групи са недостатъчно или прекомерно представени в данните за обучение.

Пристрастност при измерването



Когато използваните данни не отразяват точно това, което целят да оценят.

Пристрастие при проектирането или внедряването



Въвежда се чрез човешки избор при дефиниране на променливи, категории или критерии за оценка.

Пристрастие при потвърждаване



Когато разработчиците или дизайнерите избират данни или характеристики, които подсилват съществуващите предположения.



Рискове, свързани с алгоритмична пристрастност

Как данните могат да доведат до несправедливи резултати в образователните системи с изкуствен интелект.



Алгоритмичната пристрастност често произтича от самите данни.

Когато наборите от данни са непълни, небалансирани или лошо структурирани, системите за изкуствен интелект могат да генерират изкривени или несправедливи резултати.

Основни рискове, свързани с данните:

- **Недостатъчно представяне на определени групи ученици** → Опитът и нуждите на някои ученици може да не бъдат адекватно отразени.
- **Небалансирани набори от данни** → Системата може да работи по-добре за мнозинствените групи, отколкото за малцинствата.
- **Използване на прокси променливи (непреки показатели)** → Променливи като пощенски код или тип училище могат неволно да отразят социално-икономически или демографски различия.
- **Данни с ниско качество или несъвместими данни** → Неточната или остарялата информация може да изкриви прогнозите и препоръките.
- **Прекомерно разчитане на количествени данни** → Сложните образователни реалности могат да бъдат сведени до опростени показатели.

Ако данните са пристрастни, дори технически издържан алгоритъм може да доведе до неравностойни резултати.

Въздействието на алгоритмичната пристрастност върху равенството в образованието

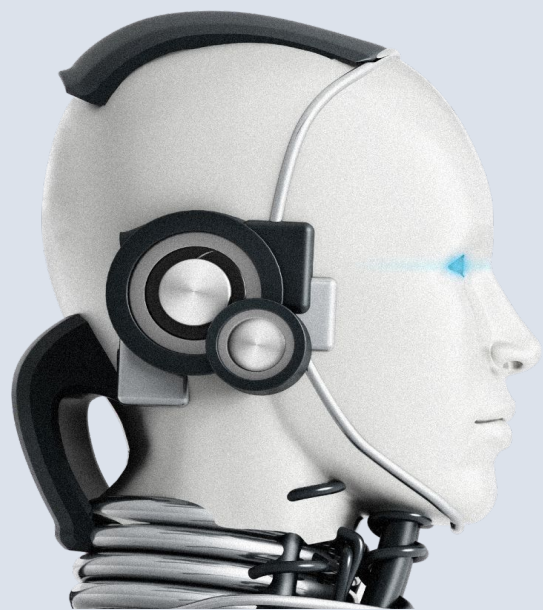
Как системите за изкуствен интелект могат да повлияят на равенството на възможностите във висшето образование.

Когато системите за изкуствен интелект съдържат пристрастия, те могат да засегнат различни групи ученици по неравностоен начин, дори и без изрично намерение. Потенциалните последствия в образователния контекст са:

- **Несправедливи или непропорционални резултати от оценяването.** Определени профили на ученици могат постоянно да получават по-неблагоприятни оценки или препоръки.
- **Засилване на вече съществуващи неравенства.** Историческите модели могат да се възпроизведат в прогнозните анализи или препоръките за академично развитие.
- **Ограничаване на академичните възможности.** Системите за насочване или мониторинг могат да повлияят на бъдещите избори и напредъка на учениците.
- **Загуба на доверие в образователните технологии.** Възприеманата несправедливост може да подкопае доверието в цифровите инструменти и институционалните практики.
- **Въздействие върху приобщаването и разнообразието.** Конкретни групи могат да бъдат систематично поставяни в неравностойно положение.

Система за прогнозиране, предназначена да идентифицира „студенти в риск“, може непропорционално да маркира студенти от неблагоприятни среди, ако историческите данни за отпадане отразяват съществуващи структурни неравенства.





Практически действия за намаляване на алгоритмичната пристрастност и насърчаване на равнопоставеността в ИИ в образованието

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

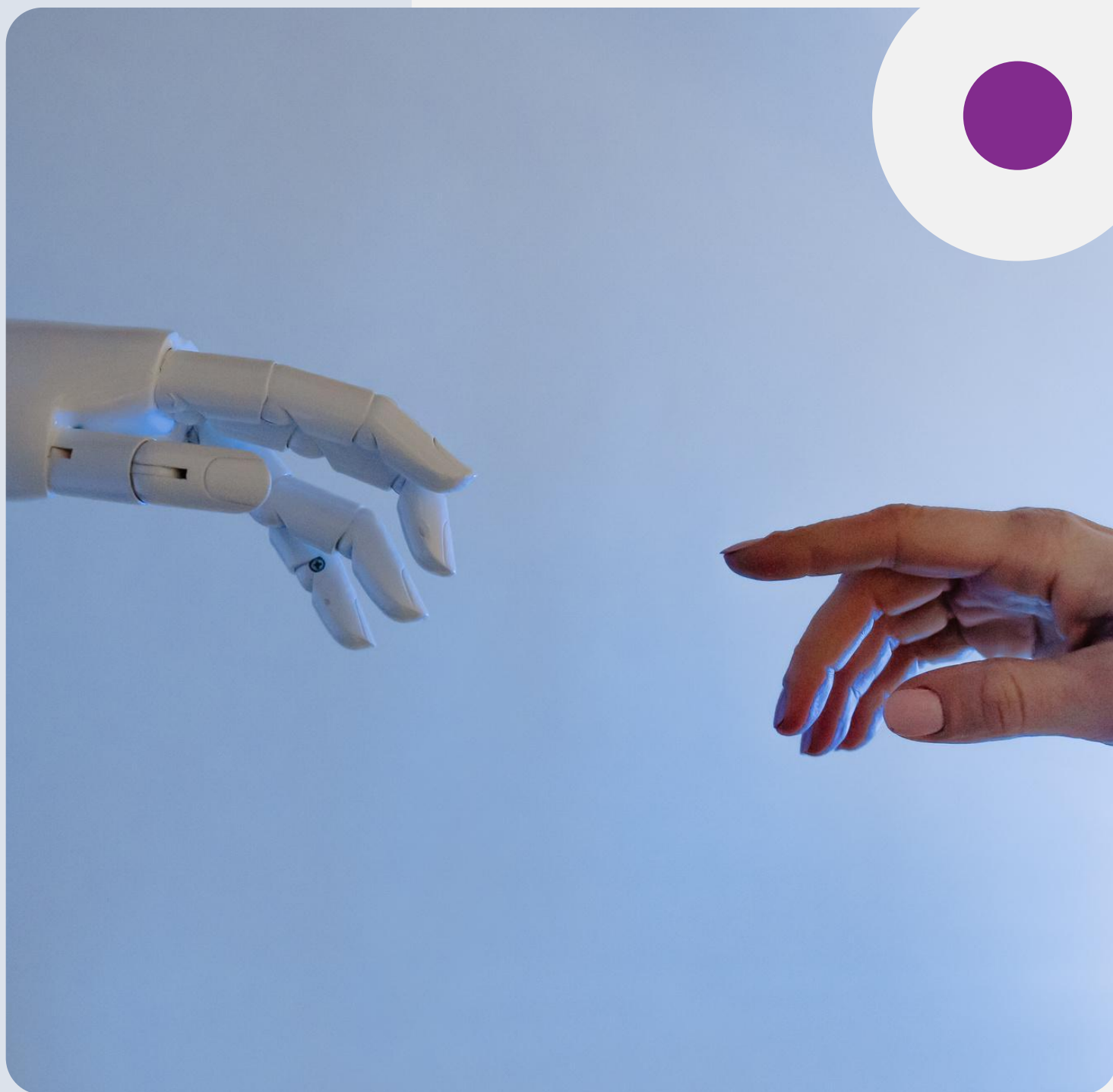
Въпреки че пристрастността не винаги може да бъде напълно елиминирана, тя може да бъде идентифицирана и намалена чрез съзнателни решения в преподавателската практика.

- Попитайте се дали използваните данни представят адекватно всички профили на учениците.
- Поддържайте активен човешки надзор и преглеждайте препоръките, генерирани от ИИ, преди да вземете академични решения.
- Насърчавайте прозрачността, като обяснявате на учениците как се използват инструментите за изкуствен интелект и какви са техните ограничения.
- Наблюдавайте дали определени групи са систематично засегнати.
- Проверявайте представителността на данните, като преценявате дали системата е обучена на разнообразни набори от данни или дали някои групи може да са недостатъчно представени.
- Сравнете дали системата дава систематично различни резултати въз основа на пол, произход, език или други релевантни фактори.
- Изслушвайте гледните точки на студентите, за да установите дали някоя група е неравно третирана.



Съфинансирано от
Европейския съюз





Можете ли да идентифицирате пристрастието?

Приложете наученото

Алгоритмичната пристрастност не винаги е очевидна. Тя често се проявява по фин начин в образователните инструменти и системите за вземане на решения. В тази дейност ще анализирате реални образователни сценарии и ще идентифицирате вида на пристрастието, което се проявява в тях.

Приложете знанията си на практика и тествайте критичния си поглед.

Започнете тук



Библиография

- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Справедливост и машинно обучение: Ограничения и възможности*.
<https://fairmlbook.org>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). Проучване относно пристрастността и справедливостта в машинно обучение. *ACM Computing Surveys*, 54(6), 1–35.
<https://doi.org/10.1145/3457607>
- ЮНЕСКО. (2021). *Препоръка относно етиката на изкуствения интелект*.
<https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- Европейска комисия. (2022). *Етични насоки за използването на изкуствен интелект и данни в преподаването и ученето за педагози*.
<https://education.ec.europa.eu/news/ethical-guidelines-on-the-use-of-artificial-intelligence-and-data-in-teaching-and-learning-for-educators>





Насърчаването на равенството в образованието изисква не само добри намерения, но и критично осъзнаване на това как технологиите формират възможностите.

Партньорството:



<https://trustai.unibit.bg/>



Проект „Trust AI“

Съфинансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Референтен номер на проекта: 2024-1-BG01-KA220-HED-000249185



Съфинансирано от
Европейския съюз

