Fairness (Férovost v AI)

Fairness v oblasti AI označuje proces zajištění toho, aby rozhodnutí učiněná modely strojového učení byla spravedlivá a nezvýhodňovala nebo nepoškozovala určité jednotlivce či skupiny. Protože se modely učí z historických dat, mohou snadno převzít a dokonce posílit lidské předsudky, které jsou v těchto datech obsaženy.

1. Zdroje nespravedlnosti (Bias)

Zaujatost (bias) se do systému může dostat v různých fázích:

- **Bias v datech (Historical Bias):** Data odrážejí existující společenské nerovnosti. Pokud například firma v minulosti najímala převážně muže, model se naučí, že „muž“ je důležitým znakem úspěšného kandidáta.
- **Reprezentační bias:** Určitá skupina je v trénovacích datech zastoupena méně (např. algoritmus na rozpoznávání obličejů trénovaný převážně na světlé pleti bude mít vyšší chybovost u lidí s tmavou pletí).
- **Algoritmický bias:** Samotná matematická optimalizace modelu může upřednostňovat většinovou skupinu, aby dosáhla co nejvyšší celkové přesnosti, i za cenu chyb u menšin.

2. Jak měřit férovost?

Existuje několik matematických definic férovosti, které si však mohou navzájem odporovat:

Metrika	Definice	Příklad
Demographic Parity	Pravděpodobnost kladného výsledku by měla být stejná pro všechny skupiny.	Stejné procento schválených půjček pro muže i ženy.
Equal Opportunity	Model by měl mít stejnou úspěšnost v identifikaci „dobrých“ kandidátů napříč skupinami.	Stejná míra (True Positive Rate) u všech etnik.
Individual Fairness	Podobní jedinci by měli dostat podobné výsledky.	Dva lidé se stejným příjmem a historií dostanou stejný úrok.

3. Metody nápravy (Mitigation Strategies)

Boje proti zaujatosti se vedou ve třech fázích životního cyklu modelu:

```
1. **Pre-processing:** Úprava trénovacích dat (např. převážení vzorků nebo odstranění citlivých atributů).
2. **In-processing:** Změna samotného algoritmu přidáním "pokuty" za nespravedlivá rozhodnutí přímo do ztrátové funkce (loss function).
3. **Post-processing:** Úprava konečných výsledků modelu tak, aby splňovaly zvolená kritéria férovosti.
```

[Image showing AI fairness intervention stages: pre-processing, in-processing, and post-processing]

4. Proč je to důležité?

- **Právní rizika:** Regulační rámce jako **EU AI Act** zavádějí přísná pravidla pro „vysoce rizikové“ AI systémy (v náboru, soudnictví, zdravotnictví).
- **Důvěra veřejnosti:** Skandály s nespravedlivými algoritmy (např. COMPAS pro předpovídání recidivy v USA) mohou vést k naprosté ztrátě důvěry v AI.
- **Ekonomický dopad:** Pokud model nespravedlivě vyloučí bonitní klienty, firma přichází o zisk.

5. Paradoxy férovosti

Jedním z největších problémů je, že **nelze vyhovět všem definicím férovosti najednou**. Často musíme obětovat malou část celkové přesnosti (accuracy) výměnou za větší spravedlnost. Tomu se říká **Fairness-Accuracy Trade-off**.

> **Zajímavost:** Amazon musel v roce 2018 zrušit svůj experimentální náborový nástroj založený na AI, protože se ukázalo, že systematicky diskriminuje ženy. I když z dat odstranili informaci o pohlaví, model se naučil identifikovat „mužský styl“ psaní životopisů a preferoval ho.

Zpět na AI rozcestník