BIAS (Algoritmická předpojatost)

Bias v IT a umělé inteligenci označuje situaci, kdy algoritmus produkuje systematicky zkreslené výsledky kvůli chybným předpokladům v procesu učení, návrhu nebo kvůli nekvalitním datům. V oblasti AI může bias vést k diskriminaci skupin lidí nebo k chybným predikcím.

1. Základní typy Biasu

Data Bias (Zkreslení v datech)

Nejčastější forma. Pokud jsou trénovací data historicky nespravedlivá nebo nereprezentativní, model se tyto předsudky naučí.

- **Příklad:** Pokud systém pro nábor zaměstnanců trénujete na datech z firmy, kde historicky pracovali jen muži, model začne automaticky znevýhodňovat ženské kandidátky.

Sampling Bias (Zkreslení výběru)

Vzniká, když data použitá k trénování neodpovídají realitě cílové populace.

- **Příklad:** Systém na rozpoznávání obličejů trénovaný převážně na lidech světlé pleti bude mít mnohem vyšší chybovost u lidí jiné barvy pleti.

Algorithmic Bias (Algoritmické zkreslení)

Vzniká chybou v samotném kódu nebo v nastavení priorit algoritmu (např. přílišná optimalizace na jeden parametr na úkor ostatních).

2. Bias-Variance Tradeoff (Statistický pohled)

Ve strojovém učení existuje základní dilema mezi dvěma typy chyb:

Pojem	Popis	Důsledek
High Bias	Model je příliš jednoduchý (Underfitting).	Ignoruje důležité vztahy v datech a dává konzistentně špatné výsledky.
High Variance	Model je příliš komplexní (Overfitting).	Příliš se soustředí na náhodný šum v trénovacích datech a selhává u nových dat.

3. Dopady v reálném světě

Bias v AI má etické i právní důsledky:

- **Soudnictví:** Algoritmy pro predikci recidivy mohou vykazovat rasovou předpojatost.
- **Finance:** Skóringové systémy bank mohou neoprávněně zamítat půjčky lidem z určitých PSČ.
- **Zdravotnictví:** Diagnostické nástroje mohou být méně přesné pro demografické skupiny, které byly v klinických studiích zastoupeny méně.

4. Jak s Biasem bojovat (Mitigace)

Boj proti předpojatosti je nekončící proces, který zahrnuje:

- **Audit dat:** Kontrola, zda jsou trénovací sady vyvážené.
- **Explainable AI (XAI):** Snaha o to, aby rozhodovací procesy „černých skříněk“ AI byly srozumitelné pro lidi.
- **Diverse Teams:** Zapojení lidí z různých prostředí do vývoje, aby identifikovali předsudky, které by programátoři mohli přehlédnout.
- **Adversarial Debiasing:** Technika, kde se jedna část AI snaží najít bias v té druhé a tím ji nutí se jej zbavit.

> **Důležité upozornění:** Úplné odstranění biasu je prakticky nemožné, protože i samotná data jsou produktem lidské společnosti, která je přirozeně subjektivní. Cílem je tedy bias identifikovat, minimalizovat a transparentně o něm informovat.

Zpět na Etiku v AI