MUNI
FI

Zkreslení v algoritmech

Mgr. Tomáš Foltýnek, Ph.D.
foltynek@fi.muni.cz

ETIKA A IT

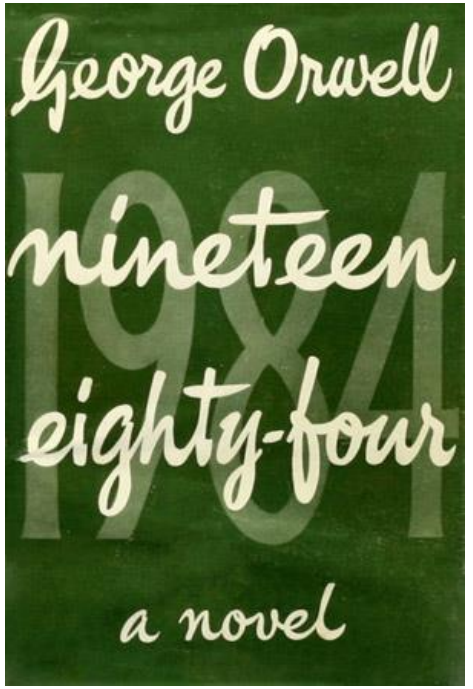
Obrázek vygenerován nástrojem Midjourney /imagine machine giving biased output

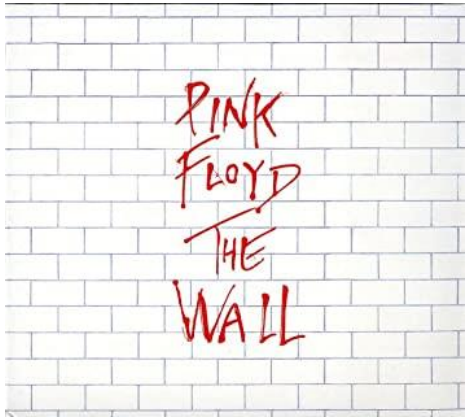
Osnova dnešní přednášky

- Opakování
  - Filtrování informací a cenzura
- Zkreslení v algoritmech
  - Obecný proces strojového rozhodování
  - Příčiny zkreslení
  - Příklady důsledků – rasová a genderová diskriminace
  - Prevence zkreslení
- Dilemma game: Rasová zaujatost

Opakování: Nová Velká čínská zeď

- Chrání zemi před ”hrozbami” jako např.
  - občanská společnost, svoboda slova
  - falešné představy a myšlenky
- Blokování ”závadného” obsahu
  - Včetně celých domén (bbc.co.uk, seznam.cz,…)
  - Filtrování výsledků hledání v Googlu
  - Uvěznění a zmizení autorů nepohodlného obsahu
- Vzniká státem kontrolovaná informační bublina
  - Mnohými (většinou společnosti?) vnímána pozitivně
  - Pomáhá udržovat současné mocenské zřízení





Opakování: EU: Právo být zapomenut

- Právo na odstranění odkazů na určité stránky **z inernetových vyhledávačů** a dalších adresářů
  - Nikoliv na odstranění obsahu jako takového
- Zohlednění vývoje osobnosti člověka
  - chyby z minulosti, které byly potrestány / napraveny / odpuštěny
  - obvinění, která byla vyvrácena
  - aktivity / události, které již nejsou relevantní
  - mohou vést k opakované stigmatizaci a znevýhodňování
- Hodnoty, které jsou v konfliktu
  - svoboda slova, právo na informace
  - právo na soukromí, ochrana osobnosti
- Google zveřejňuje aktuální data o počtech požadavků na odstranění odkazů
  - Úspěšná je cca polovina požadavků

Opakování: Blokování dezinfo webů

- Rekonstrukce státu spolu s dalšími organizacemi vydali Doporučení pro blokování serverů
  - Kritérium konkrétní, bezprostřední a vážné újmy
    - Jasná definice, kdy lze k vypnutí přistoupit
  - Využití širší sady nástrojů
    - Výzva provozovateli, stažení jednotlivého článku,…
  - Soulad s lidskoprávní judikaturou
    - Ochrana demokracie před hrozbami je dvojsečná
  - Řádný proces včetně možnosti odvolání
    - Lhůty, zveřejněné zdůvodnění
  - Pravomoc
    - Mezioborový úřad, soud
  - Posilování role občanské společnosti
    - Vyvracení dezinformací, šíření pravdy

Zkreslení v algoritmech

Zdroj obrázku: https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency

Úvod

- Informační společnost stále více využívá algoritmy
- Algoritmy mohou
  - Upozornit
  - Poradit
  - Rozhodnout
- Algoritmy určují, jak interpretovat data
  - Tedy jak transformovat data na informace
- Cílem algoritmu je optimalizace procesu
  - Vůči určitému kritériu

Příklady (eticky relevantní)

- Doporučující systémy pro vyhledávání
  - Pořadí webových stránek / videí / písniček / odborných článků / zboží
  - Aneb proč je AAA Auto nejúspěšnější autobazar?
- Navigace
  - Optimální trasa – Nejkratší? Nejrychlejší? Nejlevnější?
  - S ohledem na řidiče, nebo s ohledem na místní obyvatele?
- Sociální média
  - Které příspěvky / reklamy zobrazit
- Investiční strategie
  - Jaké akcie koupit / prodat
- Analýza rizik
  - Půjčit či nepůjčit danému klientovi?
  - S jakým úrokem?
- Rozpoznávání objektů ve scéně
  - Je chodec v cestě automobilu?
  - Za jakou cenu se mu vyhnout?

Obecný proces

Data → Informace → Rozhodnutí → Akce → Etické důsledky

- Optimalizace na daná kritéria ≠ eticky přijatelné rozhodování
- Strojové učení → Nepředvídatelnost

- Kvalita rozhodování závisí na
  - Kvalitě dat (”garbage in, garbage out”)
  - Kvalitě procesu, který je zpracovává

- Proces nemusí být možné zrekonstruovat / vysvětlit / obhájit
- Podstatná je **akce** a její **etické důsledky**

Netransparentnost rozhodování

- “Když to říká počítač, musí to být pravda”

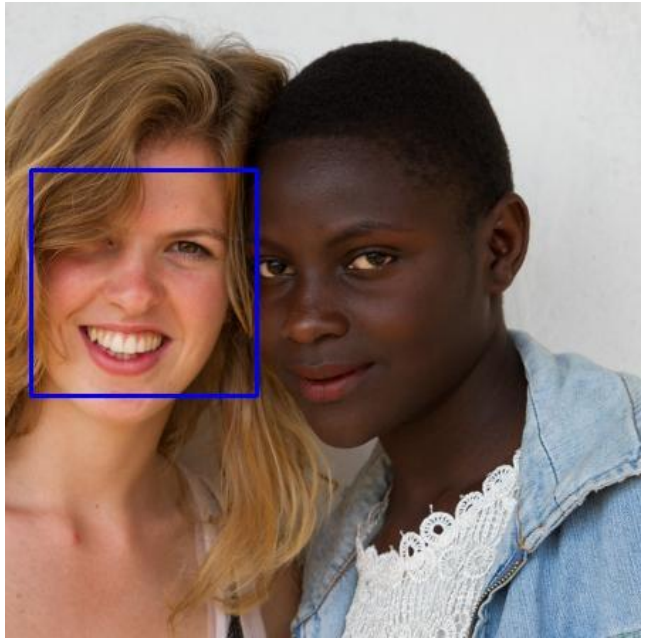
- Rozhodování založené na korelaci, nikoliv kauzalitě
- Netransparentní, proprietární algoritmy
  - Konkurenční výhoda
  - Bezpečnost
  - Prevence záměrných manipulací (“gaming the system”)
- Black box problem
  - Nemožnost obhájit rozhodnutí → Ztráta důvěry v systém

Příčiny zkreslení

- Umělá inteligence se vždy učí na **starých** datech
  - Ta mohou být zkreslená
- Odlišné hodnoty při návrhu systému
  - Manuální “tagování” zohledňuje hodnoty tagujících
  - Nastavení ”ground truth” pro učení klasifikátorů
- Technická omezení
  - Zjednodušení algoritmu
  - Využití externích knihoven (včetně zkreslení)
- Jiný kontext užití

Příklad: Proctorio

- Software pro sledování studentů během online testů
  - Sledování aktivity – student používá pouze prohlížeč s testem
  - Sledování obličeje studenta – identifikace, detekce “podezřelého" chování
- Rozpoznávání obličejů pomocí OpenCV
  - Úspěšnost u černochů <50%
- Studenti tmavší pleti častěji označováni jako podezřelí



- Více informací (a zdroj obrázku)

Identifikace obličejů

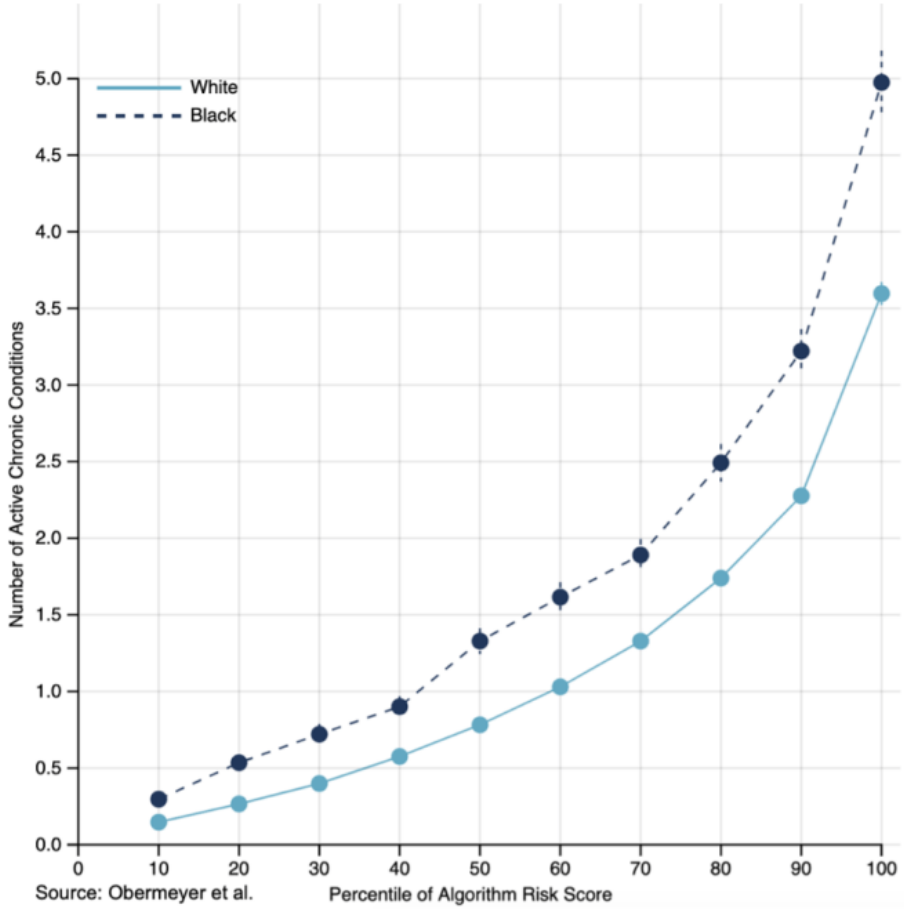
- Nižší úspěšnost není problém jen OpenCV
- Všechny systémy vykazují nižší úspěšnost u černochů a u žen
  - Profesionální systémy pro porovnávání fotografií (pasová kontrola)
  - Chybovost u bílých mužů cca 1 : 10 000
  - U černých žen je chybovost 5x – 10x vyšší
  - Obojí je stále násobně lepší než vyhodnocování člověkem
- Důvody – “pale male data”
  - Větší zastoupení bělochů a mužů v trénovacích datech (obrázky stažené z webu)
  - Makeup u žen
  - Větší kvalita obrázků bělochů
    - Optimalizace fotoaparátů na bílé obličeje + kvalitnější vybavení

Příklad: Zdravotní péče v USA

- Prevence je levnější než léčba
- Umělá inteligence identifikuje pacienty, kteří by měli dostat preventivní péči

- Ideální cíl
  - Čím větší **očekávaná nemocnost**, tím vyšší priorita pro preventivní péči
- Implementovaný cíl
  - Čím větší **očekávané výdaje na zdravotní péči**, tím vyšší priorita pro preventivní péči

- Problém: Diskriminace černochů
  - Horší péče → Levnější léčba → Nižší priorita



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Personalizace → Diskriminace

- Personalizace → Segmentace populace
- Jen určité segmenty mají nárok na určité služby / informace → Prohlubování rozdílů mezi segmenty

- Personalizace cenových nabídek

Příklad: Amazon a nabírání zaměstnanců

- Amazon – snaha o automatizaci úplně všeho
  - Jeden z důvodů komerčního úspěchu
- Automatické filtrování CV při náboru nových zaměstnanců
  - Snaha ušetřit práci HR oddělení
  - Vstup: Velké množství životopisů a motivačních dopisů
  - Výstup: Jména *n* nejlepších kandidátů
- Systém se učil na životopisech již nabraných zaměstnanců
  - To byli především muži
- Důsledek: Zvýhodňování mužů oproti ženám
- Vývoj systému zastaven před jeho uvedením do provozu (2018)

Systémy pro posuzování CV

- Odhad: ¾ životopisů v USA přečte algoritmus, aby rozhodl, jestli je vůbec ukázat člověku
- Zvýhodňující znaky
  - Thomas, church
  - Jarred, lacross
- Znevýhodňující znaky
  - women's college, women's chess/socker/… club
  - mezera mezi zaměstnáními delší než 6 měsíců
- Firmy uvádějí stále více požadavků
  - což dále odrazuje ženy (muži se hlásí, i když požadavky nesplňují)

Zdroje: Video, článek

Prevence diskriminace

- Návrh aplikace
  - Nediskriminující optimalizační kritéria
- Kontrola trénovacích dat
  - Statistické testy vůči různým skupinám populace
- Začlenění antidiskriminačních kritérií do klasifikačních algoritmů
  - Záměrné zvýhodňování diskriminovaných skupin (?)
- Ex-post kontrola férovosti rozhodování
  - Statistické testy vůči různým skupinám populace

Kdo nese zodpovědnost?

- Tradiční pojetí
  - Programátor rozumí veškerému kódu
  - Je zodpovědný za jeho fungování
  - Kód je výsledkem promyšleného návrhu
  - Rozhodovací mechanismy jsou součástí kódu
- Dnešní realita “černých skříněk”
  - Externí knihovny
  - Strojové učení
- Mezera v zodpovědnosti
  - Co má návrhář / vývojář pod kontrolou vs. jak se algoritmus chová

Regulace AI

- UNESCO Recommendation on the Ethics of AI
  - Výzva k dobrovolnému začlenění do legislativy členských států
  - Definuje hodnoty a principy, stanovuje oblasti aplikace
- The Partnership on AI to Benefit People and Society
  - Nezisková organizace založená velkými firmami
  - Zdroje, doporučení, příklady dobré praxe
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
  - Guidelines, standardy,…
- EU AI Act

EU AI Act

- První právně závazný předpis regulující používání AI
- Čtyři úrovně rizika

Úroveň rizika	Oblast použití	Požadavky
Neakceptovatelné	Sociální skóre	Úplný zákaz
Vysoké	Kritická infrastruktura, vzdělávání, bezpečnostní součástky, trh práce, finanční služby, právo, migrace,…	Systém řízení rizik Kvalita trénovacích dat Logování aktivit Podrobná dokumentace Jasné informace pro uživatele Dohled člověka Robustnost, bezpečnost, přesnost
Omezené	Chatboti	Transparentnost
Nízké nebo žádné	SPAM filtry, počítačové hry	Žádná regulace

Závěrem

- Je třeba rozlišovat
  - Chyby v návrhu vs. chyby vzniklé za běhu
  - Nezamýšlené vedlejší efekty vs. úmyslné zkreslení

- I zkreslené algoritmy jsou mnohdy lepší než lidé
- Odstranit zkreslení algoritmu je obvykle jednodušší než odstranit předsudky člověka

- Je třeba definovat **ideální cíl** a zajistit, že
  - tento cíl je etický
  - algoritmy tento cíl skutečně sledují

Dilemma Game

Obrázky generovány nástrojem Midjourney /imagine people arguing ethical dilemmas

Dilema: Rasová zaujatost

Pracuji jako senior vývojář algoritmů strojového učení. Aktuálně pracujeme na algoritmu, který má pomoci snížit následky kriminality pomocí automatizovaného posuzovaní nebezpečnosti chování z kamerového záznamu a vyslání hlídky.

Algoritmus v testovací fázi funguje velmi dobře a policie si jej pochvaluje. Zjistili jsme ale, že mezi falešnými poplachy jsou výrazně častěji občané černé pleti. Mám možnost do algoritmu zasáhnout tak, aby barvu pleti zohledňoval méně, to ovšem s velkou pravděpodobností sníží úspěšnost celého programu. Co bych měl udělat?

A. Nechám všechno tak jak je a budu doufat, že algoritmus se sám naučí, že na barvě pleti záleží méně než na jiných faktorech, a tím se sníží počet falešných poplachů.

B. Algoritmus výrazně upravím a zcela mu zakážu brát barvu pleti jako faktor. Namísto toho se pokusím přeprogramovat ho tak, aby se zaměřil na jiné indikátory navzdory tomu, že tato snaha nemusí být úspěšná.

C. Kvůli obavě z nařčení z rasové diskriminace a nesnášenlivosti navrhnu pozastavení, případně ukončení vývoje programu z důvodu nedokonalosti.

D. Budu si stát za svým algoritmem, přičemž budu argumentovat, že statisticky barva pleti bývá zásadní při určovaní potenciální hrozby.

Příští týden – Mgr. Jan Kvapil

- Téma: Etický hacking
  - Jak funguje mozek hackera
  - Jak hackovat legálně (bug bounty programy)
  - Proč kvůli výzkumu na FI museli všichni Estonci měnit občanky

- Úkoly
  - Poslechnout si podcast z Darknet Diaries o hacknutém twitterovém účtu Donalda Trumpa: https://darknetdiaries.com/episode/87/