

MUNI  
FI

# Zkreslení v algoritmech

Mgr. Tomáš Foltýnek, Ph.D.

[foltynek@fi.muni.cz](mailto:foltynek@fi.muni.cz)

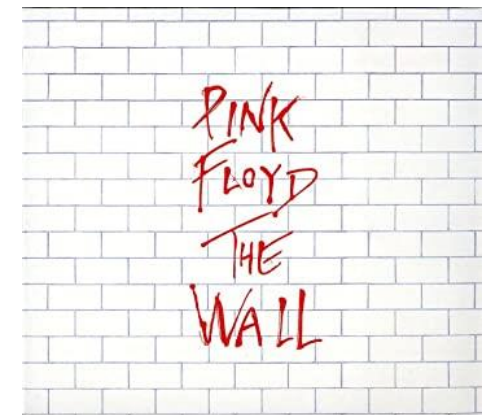
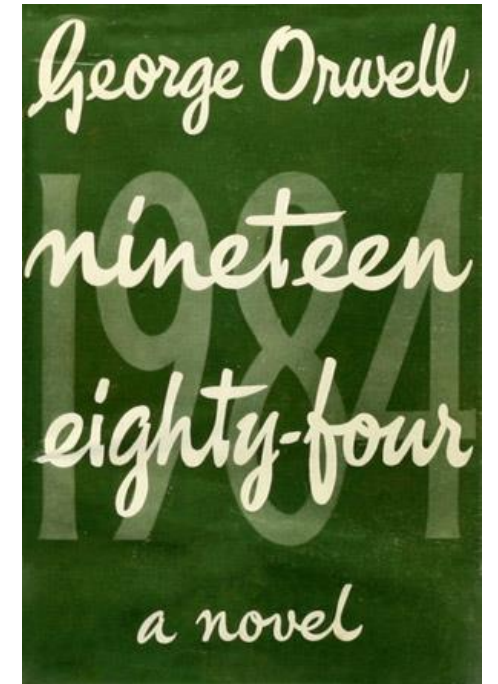


# Osnova dnešní přednášky

- Opakování
  - Filtrování informací a cenzura
- Dilemma game: Filtrování v univerzitní síti
- Martin Tuček: COMPAS
- 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 zed'

- 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 společností?) vnímána pozitivně
  - Pomáhá udržovat současné mocenské zřízení



# Opakování: EU: Právo být zapomenut

- Právo na odstranění odkazů na určité stránky **z internetových 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

- ČR připravuje zákon na boj s dezinformacemi (SR už má)
- Rekonstrukce státu spolu s dalšími organizacemi vydali [Doporučení pro vypínání 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

# Filtrování v univerzitní síti?

- Problém: Contract cheating
  - Studenti si najmou někoho, kdo za ně vypracuje úkol
    - Typicky seminární či závěrečná práce
  - Je to forma plagiátorství
    - Studenti odevzdají cizí dílo, které vydávají za své vlastní
    - Nejistitelné současnými antiplagiátorskými systémy
  - Problém zejména v anglosaském vzdělávacím systému
    - Znamky ve velké míře závisejí na samostatných úkolech
    - 3-8 % student se přiznává, že se někdy během studia contract cheatingu dopustilo
    - 2/3 to dělají opakovaně → Pro 2-5 % studentů je to standardní modus operandi
- Mnoho firem inzeruje své služby jako „pomoc s psaním práce“
  - Studenti se nechávají nalákat na zdánlivě legální službu
  - Reklamní kampaně jsou často poměrně agresivní a přesvědčivé

# Filtrování v univerzitní síti?

- Některé univerzity blokují přístup na tyto stránky z vnitřní sítě
  - Doporučení např. britské QAA
- Jiné jen zobrazí varování, ale uživatele pustí dále
- Obdobný problém: Chat GPT
  - Některé školy blokují přístup
- Je filtrování správné?
  - Jaká jsou / mají být kritéria?
  - Kdo má rozhodovat o zařazení na seznam blokováných stránek?



The screenshot shows a web filtering error page from Massey University. The header includes the Massey University logo and the text 'Web Filtering' and 'Powered by ITS Security'. The main content features a red 'X' icon and the heading 'Web Page Blocked!'. The text explains that the website is blocked due to inappropriate content and provides contact information for the ITS Service Desk. Technical details such as the URL, category, client IP, server IP, user name, and group name are listed. A note at the bottom states that access to the site requires authorization from a department head or manager.

MASSEY UNIVERSITY  
UNIVERSITY OF NEW ZEALAND

Web Filtering

Powered by ITS Security

**Web Page Blocked!**

The website you have requested has been deemed as inappropriate. This activity breaches Massey University's Internet access policies. The website has been blocked due to clear, historical breaches of policy.

Site categorisation is not 100% accurate, so if you feel you have received this message in error, please contact the ITS Service Desk via the following [online help request form](#), or by phone on extension 82111.

Internet usage is routinely monitored and logged.

URL: proxyturbo.com/  
Category: Proxy Avoidance  
Client IP: 130.123.32.165  
Server IP: 107.167.86.174  
User name:  
Group name:

If you have a legitimate study, research or work related reason to access this site, or if you have questions regarding any of the information in this notice, please contact the ITS Service Desk via the following [online help request form](#), or by phone on extension 82111.

Note that permission for this access will have to be authorised by your Head of Department or your manager. Massey University's Internet usage policies for staff and students can be found [here](#).

# Dilemma Game





# Dilema: Filtrování v univerzitní síti

Mnohé webové stránky představují hrozbu pro návštěvníky. Jedná se o bezpečnostní hrozby (phishing, malware), nelegální obsah, dezinformace, etická rizika, atd. Jak by měla univerzita přistupovat k filtrování těchto hrozeb?

- A. Nijak. Svoboda slova je nade vše, na univerzitu žádné filtrování nepatří.
- B. Filtrovat pouze prokazatelné bezpečnostní hrozby, vše ostatní by mělo být povoleno.
- C. Svoboda slova má své hranice. Univerzita může v závažných a odůvodněných případech výjimečně přistoupit k blokování vybraných stránek či domén.
- D. Univerzita je vzdělávací instituce. Je v pořádku (rutinně) blokovat jakékoliv nevhodné stránky, které uživatelé nepotřebují ke svojí práci nebo studiu.

# Zkreslení v algoritmech



# 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

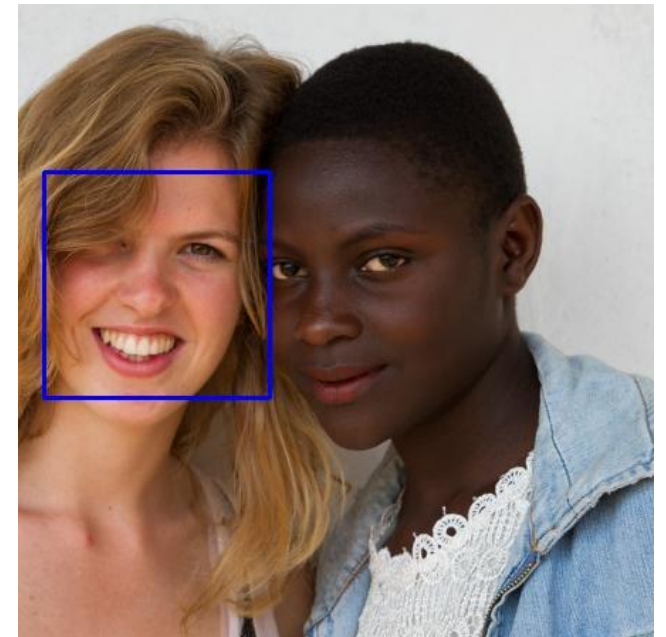
- “Když to říká počítač, musí to být pravda”
- Rozhodování založené na korelaci, nikoliv kauzalitě
- Netransparentní, proprietární algoritmy
  - Konkurenční výhoda
  - Bezpečnost
  - Prevence záměrných manipulací (“gaming the system”)
- Black box problem
  - Nemožnost obhájit rozhodnutí → Ztráta důvěry 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 černochoů <50%
- Studenti tmavší pleti častěji označováni jako podezřelí
- [Více informací \(a zdroj obrázku\)](#)



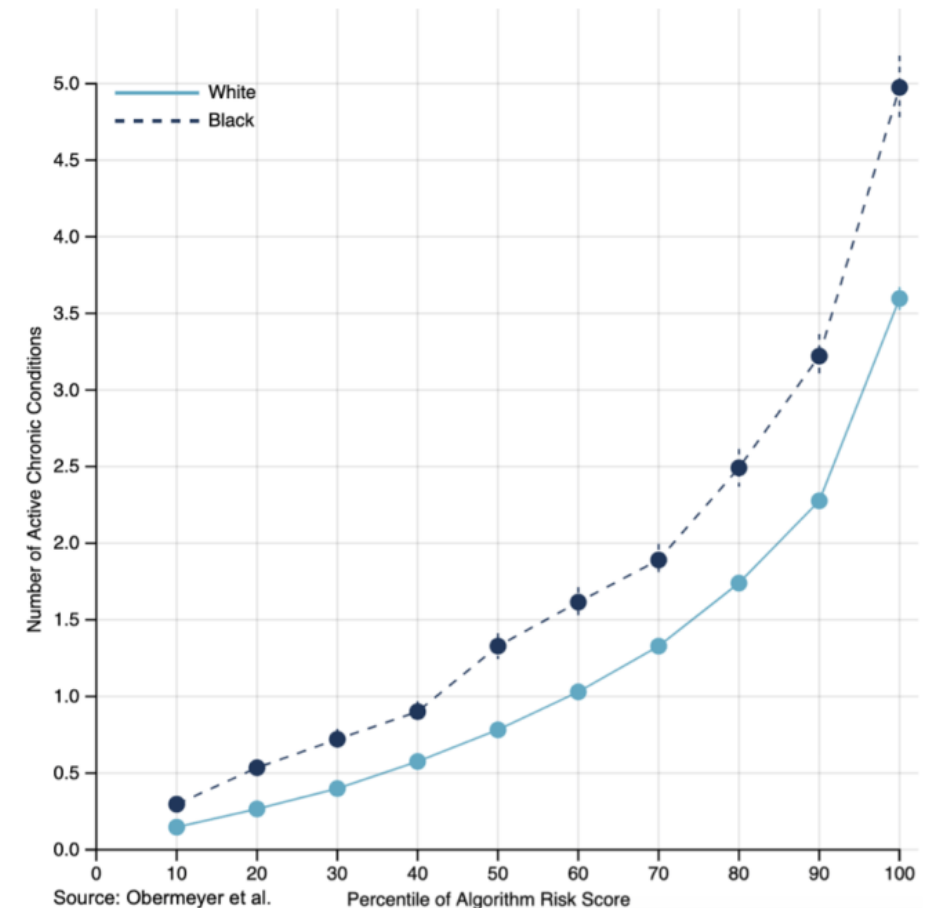


# Identifikace obličejů

- Nižší úspěšnost není problém jen OpenCV
- Všechny systémy vykazují nižší úspěšnost u černochoů 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 černochoů
  - 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 naboraných zaměstnanců
  - To byli především muži
- Důsledek: Zvýhodňování mužů oproti ženám
- Vývoj systému zastaven před jeho uvedením do provozu (2018)

# Systemy pro posuzování CV

- Odhad:  $\frac{3}{4}$  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

# EU AI Act (v přípravě)

- Reguluje jakékoliv využívání nástrojů založených na AI
  - Měl by se stát de facto světovým standardem jako GDPR
- Tři úrovně rizika
  - Nepříjemné (např. čínské sociální skóre) → zakázáno
  - Vysoké (kritická infrastruktura, vymáhání práva, vzdělávání,...) → předmět regulace
  - Nízké (všechno ostatní) → nereguluje se
- Povinnosti provozovatelů vysoce rizikových aplikací
  - Pravidelná analýza rizik
  - Zveřejnění trénovacích dat (auditovatelnost)
- General Purpose AI Systems (GPAIS)
  - Aplikace využitelné k různým účelům (např. Chat GPT)
  - Zatím není jisté, zda budou v kategorii „vysoké riziko“
- Předpokládaný časový horizont
  - Schválení zákona během 2023
  - Účinnost za další 2 roky

Zdroj: <https://www.reuters.com/technology/what-is-european-union-ai-act-2023-03-22/>



# 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



# Dilema: Rasová zaujatost

Pracuji jako senior vývojář algoritmů strojového učení. Aktuálně pracujeme na algoritmu, který má pomoci snížit následky kriminality pomocí automatizovaného posuzování nebezpečnosti chování z kamerového záznamu a vyslaní 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. Pokusím se algoritmus ovlivnit tak, aby trochu méně řešil barvu pleti. Udělám to ale takovým způsobem, abych příliš nezhoršil celkovou úspěšnost programu.
- C. 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.
- D. Kvůli obavě z nařčení z rasové diskriminace a nesnášenlivosti navrhnu vládě pozastavení, případně zrušení programu z důvodu nedokonalosti.
- E. Budu si stát za svým algoritmem, přičemž budu argumentovat, že statisticky barva pleti bývá zásadní při určování potenciální hrozby.

# Příští týden

- Téma: Internet a autorská práva
  - Intelektuální vlastnictví a autorská práva
  - „Remixová“ kultura
  - Volné dílo a licence Creative commons
- Prezentace Martina Tvarožka:
  - Chat GPT a Stable Diffusion – soudní procesy
- Úkoly
  - Přečíst si [shrnutí knihy](#) "Made with Creative Commons"
    - Nebo [celou knihu](#) 😊
  - Zběžně se podívat na <https://creativecommons.org/>