MUNI
FI

Mediální manipulace a dezinformace na internetu

Mgr. Tomáš Foltýnek, Ph.D.

foltynek@fi.muni.cz

Osnova dnešní přednášky

- Reflexe filmu The Social Dilemma
- Mysl
- Opakování: Zkreslení v algoritmech
- Mediální manipulace a dezinformace na internetu
  - Terminologie, nejčastější zdroje informací
  - Ekonomika zpravodajských serverů
  - Šíření dezinformací na sociálních sítích
  - Detekce dezinformací
- Dilema: Kontrola dat sesbíraných pro závěrečnou práci

Reflexe filmu The Social Dilemma

- Nejčastěji zmiňované myšlenky
- Příčiny problému
  - Snaha o maximalizaci zisku
  - Snaha o maximalizaci času stráveného u obrazovky
  - Původně dobrý záměr, který se ale časem „zvrtnul“
- Podstata problému
  - Návykovost, závislost
  - My jsme produkt
- Řešení problému
  - Regulace

Opakování: Zkreslení v algoritmech

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**

Opakování: Netransparentnost

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

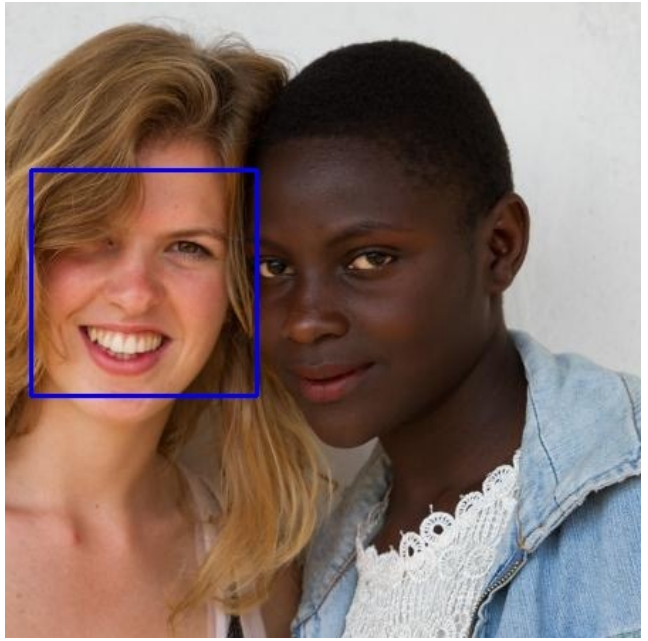
- Rozhodování založené na korelaci, nikoliv kauzalitě
- Netransparentní, proprietární algoritmy
  - Konkurenční výhoda
  - Bezpečnost
  - Prevence záměrných manipulací (“gaming the system”)
- Nemožnost obhájit rozhodnutí → Ztráta důvěry v systém

Opakování: 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

Opakování: 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)

Opakování: Identifikace obličejů

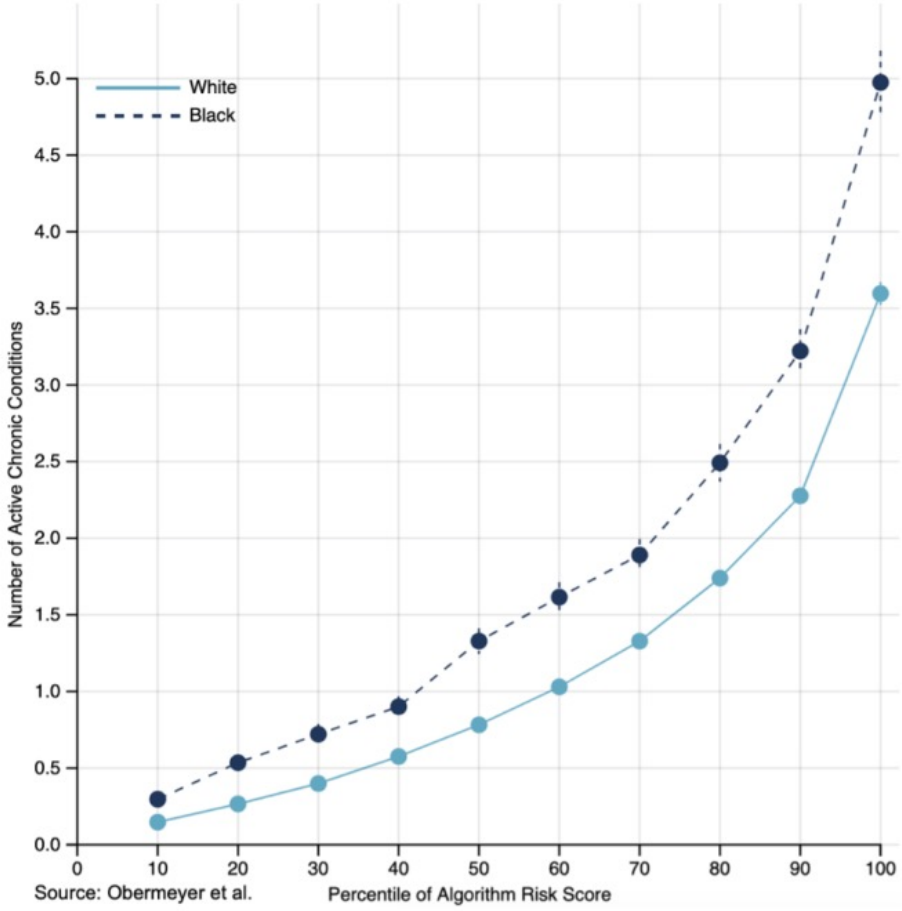
- Nižší úspěšnost není problém jen OpenCV
- Všechny systémy vykazují nižší úspěšnost u černochů a u žen
  - Profesionální systémy pro porovnávání fotografií (pasová kontrola)
  - Chybovost u bílých mužů cca 1 : 10 000
  - U černých žen je chybovost 5x – 10x vyšší
  - Obojí je stále násobně lepší než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

Opakování: 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.

Opakování: Amazon a nabírání zaměstnanců

- Amazon – snaha o automatizaci úplně všeho
  - Jeden z důvodů komerčního úspěchu
- Automatické filtrování 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

Opakování: 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

Opakování: 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

Opakování: Zkreslení v algoritmech

- 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

Opakování: Zkreslení v algoritmech

- Dotazy?
- Postřehy?
- Novinky?
- Diskuse?

Mediální manipulace a dezinformace na internetu

Zdroj obrázku: https://alphanews.org/announcing-alpha-news-media-bias-contest/

Odpovědník

- Článek ‘Nothing on this page is real’: How lies become truth in online America
  - https://www.washingtonpost.com/national/nothing-on-this-page-is-real-how-lies-become-truth-in-online-america/2018/11/17/edd44cc8-e85a-11e8-bbdb-72fdbf9d4fed_story.html
- Christopher Blair podle mého osobního názoru (studenta - respondenta odpovědníku :-))
  - dělá správnou věc
  - páchá více škody než užitku
- Uvěřili jste někdy dezinformaci? A šířili jste ji?

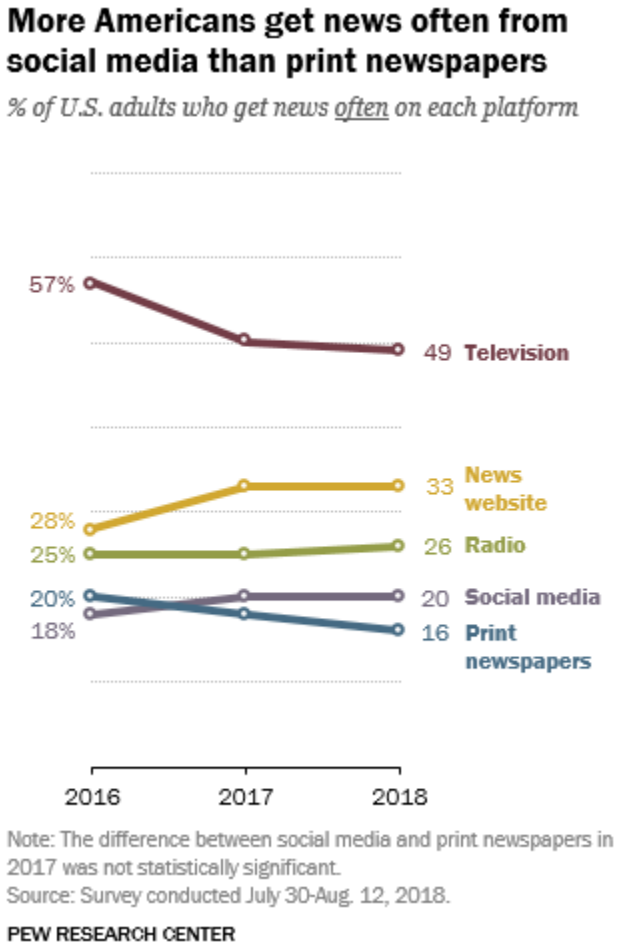
Terminologie

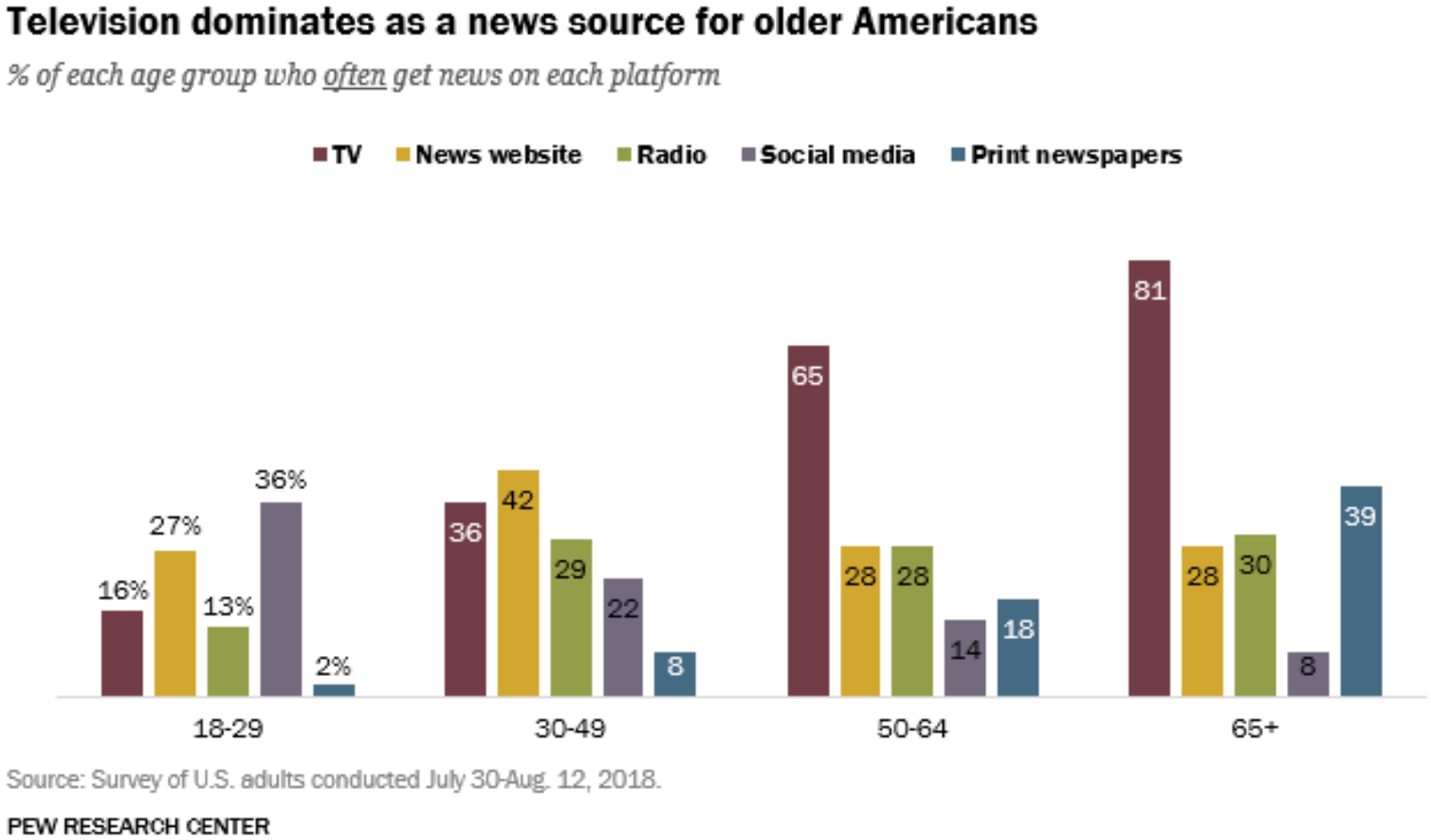
- Dezinformace
  - **Záměrně falešná informace** s cílem ovlivnit názory lidí
  - Vytvořena tak, aby **vyvolávala zdání pravdivosti**
  - Vzniká manipulací existující (pravdivé) zprávy nebo vytvořením zcela nové zprávy
- Misinformace
  - Neúmyslně chybná zpráva
  - Satira
  - Tj. bez úmyslu ovlivnit něčí názor

Psychologická podstata

- Proč lidé věří dezinformacím?
  - **Naivní realismus**: Lidé věří, že jejich vnímání reality je jediné správné
  - Potvrzovací zkreslení (**confirmation bias**): Lidé preferují informace, které potvrzují jejich vidění světa
- Je velmi obtížné důvěru narušit
  - Oprava mylných údajů (pokud pochází “z opačného tábora”) může dokonce ještě více pokřivit vnímání reality
- Lidé se snaží maximalizovat osobní užitek
  - Včetně společenského přijetí a ocenění
  - Co s podivným příspěvkem od kamaráda?
    - Lajkovat a sdílet, nebo ověřovat a vyvracet?
    - Která volba je “společensky bezpečnější”?

Zdroje informací

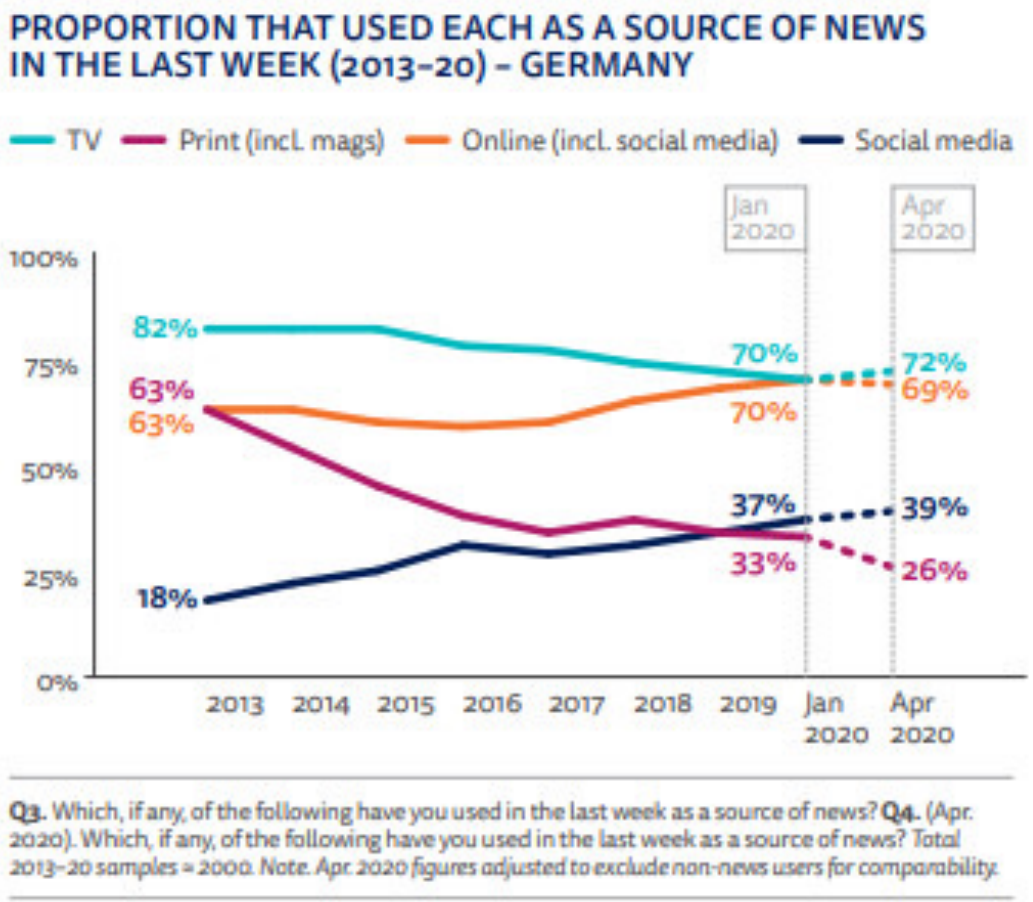




Zdroj: https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/

Zdroje informací

- Pokles TV a tištěných novin
- Vzestup online zpráv a sociálních médií
  - Sociální bubliny
  - Více náchylné na šíření dezinformací



Zpravodajské servery

Zpravodajské servery

- Zamýšlený cíl: Co nejlépe informovat o tom, co se děje
  - Přiměřeně, zodpovědně
- Realita: Optimalizace na analytiku
  - Počet zobrazení / kliků / lajků
- Většina čtenářů kliká na negativní a senzační zprávy
  - Poskytování takovýchto zpráv uspokojuje poptávku
- Vyhrávají nejsenzačnější titulky
  - Útočící na city a pudy čtenářů
- Vytváří podhoubí pro extrémní a manipulativní informace

Ekonomika zpravodajských serverů

- Mnoho autorů, omezené zdroje
- Autoři musí naplnit kvóty na délku obsahu a čtenost
  - Kvalita je druhořadá
- Investigativní žurnalistika je drahá
  - Rozhovory, ověřování informací
- Poctivou investigativní žurnalistiku čte málo lidí
- Tlak na zisk → Tlak na čtenost zpráv
- Tlak na tvorbu „levných“ zpráv
  - Lze napsat od stolu s využitím internetu
  - Místo ověřování stačí odkaz nebo screenshot
    - důvěryhodnost řeší málokdo
- Pokrytí témat jenom proto, že už jsou pokryta jinde
  - Pokud je tato zpráva čtená u konkurence, je potřeba o ní také psát

Nestrannost

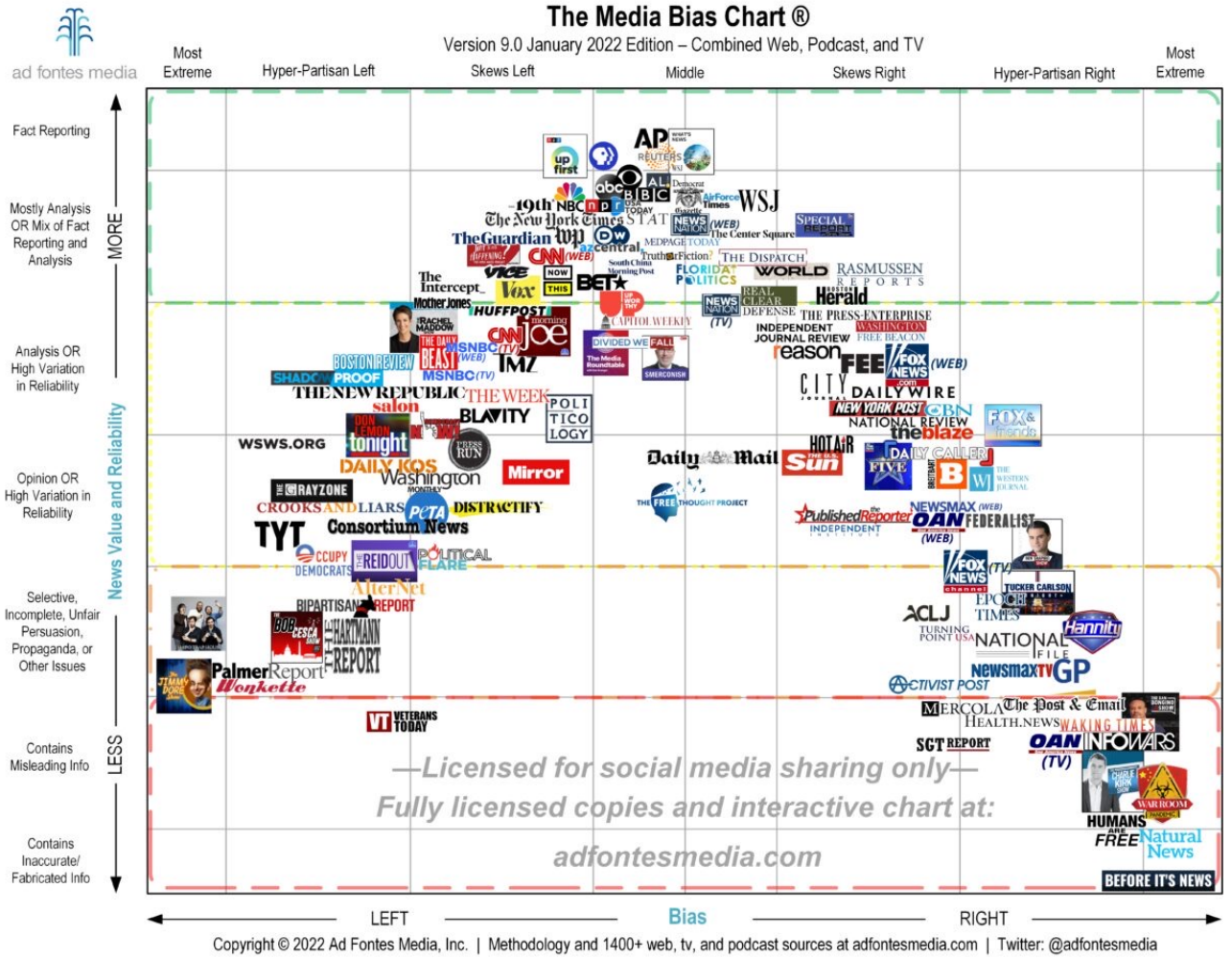
- Extrémní pojetí nestrannosti
  - Oba (všechny) protinázory dostávají stejný prostor
- Včetně nepravdivých, manipulativních, či nehumánních
- Vede k šíření názorů, které vůbec nestojí za pozornost
  - Legitimizace extremismu

- Svoboda slova jako argument pro šíření rasismu a ponižování žen
  - Rozlišení, zda něco **mohu** říct vs. zda **bych měl**
  - Obava z obvinění z cenzury
- Přitom cílem zpravodajství je předávat a vysvětlovat **fakta**

Maximalizace užitku

- Autor zprávy
  - Krátkodobý užitek: Dosah na co nejvíce čtenářů
  - Dlouhodobý užitek: Autenticita zpráv → Reputace
- Čtenář zprávy:
  - Informační užitek: Dozví se pravdivou a nepokřivenou informaci
  - Psychologický užitek: Dozví se informaci, která zapadá do jeho názorů a potřeb
- Dezinformace se šíří pokud
  - Autoři upřednostňují krátkodobý užitek
  - Čtenáři upřednostňují psychologický užitek

Media Bias Charts



https://www.allsides.com/media-bias/media-bias-chart
https://adfontesmedia.com/static-mbc/

Metody mediální manipulace

- Ředění informací
  - A odvedení pozornosti jinam
- Zdůrazňování nepodstatných problémů
  - K odvedení pozornosti od reálných (vážných problémů)
- Vytvoření umělého problému
  - A nabídka jeho řešení
- Využívání citů a pudů
  - Zejména strach (z neznámého)
- Útok na elity
  - Každý přece “selským rozumem” pozná…

Manipulace volbou slov

- Practicing pro-life litigators know that Trump judges are saving lives by permitting restrictions on abortion to go into effect.

  https://thefederalist.com/2020/04/24/david-french-needs-to-stop-slandering-trump-supporting-christians/

- Bojovník za svobodu vs. Terorista
- Uprchlíci / Migranti / Přistěhovalci
- Válka / Invaze / Agrese / Speciální vojenská operace

BREITBART

TRENDING: UKRAINE WAR | RECORD HIGH GAS | THE REOPENING | BIDEN CRIME WAVE | WOKE WAR

REPORT: OVER 80 MILLION CHILDREN AT RISK AS CORONAVIRUS DISRUPTS VACCINATION SCHEDULES

439 EMAIL PARLER TWEET

John Amis/AP

by JOSHUA CAPLAN | 22 May 2020 970

LISTEN TO STORY 5:19

Tens of millions of children under 12 months are potentially at risk for diseases such as diphtheria and polio as the Chinese coronavirus pandemic interrupts routine vaccinations, according to data published by global public health experts on Friday.

Dezinformace na sociálních médiích

Dezinformace na sociálních médiích

- Sociální boti (social bots)
  - Programy, které automaticky publikují obsah a interagují s uživateli
  - Odhadu: V den voleb Trump vs. Clinton tweetovalo cca 19 milionů botů
- Trollové
  - Účastníci online diskusí
  - Snaží se záměrně vyvolávat hádky, provokovat, urážet skupiny či jednotlivce, šířit propagandu nebo falešné informace
- Sociální kyborgové
  - Kombinace lidského vstupu a automatických programů na zvýšení dosahu
- Sociální bubliny → Šíření informací, kterým bublina věří → Posilování důvěry → Větší uzavřenost bubliny

Příklad: PizzaGate

- Konspirační teorie šířená před prezidentskými volbami v USA
- Uniklé e-maily Johna Podesty, šéfa kampaně Hillary Clinton údajně obsahovaly kódované zprávy
  - Představitelé Demokratické strany a několik restaurací v USA
  - Měly provozovat obchod s lidmi a nabízet sex s dětmi
- Šíření konspirace přes Twitter, 4chan, 8chan…
- Zaměstnanci dotčených restaurací dostávali výhružky smrtí
  - V několika případech se v restauracích i střílelo
- Konspiraci uvěřilo 46 % voličů Donalda Trumpa a 17 % voličů Hillary Clinton
- Více viz https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory

Dilemata

- Novináři: Psát o tom, jaké zprávy se šíří?
- Uživatelé sociálních sítí: Sdílet?

- Psát, sdílet?
  - A zároveň kritizovat, vyvracet, vzdělávat
  - Ale šířit původní zprávy mezi více lidí

- Ignorovat
  - Tvářit se, že neexistují
  - Neposkytnout oponentní názor
  - Čekat na něco horšího, co už nebude možné ignorovat

- Co je správné?

Šíření falešných zpráv

- Studie MIT na datech z Twitteru z let 2013-2017
- Publikovaná 2018 v Science
- Falešné zprávy se šíří 3x rychleji
- Nešíří je boti, ale skuteční lidé
  - Někteří záměrně, jiní nevědomky
- Falešné zprávy → Překvapující, nové → Chci se podělit

“Going Viral”

- Výzkum na Stanfordu 2019
- Analogie šíření falešných zpráv a šíření viru
  - Nemocný = věří zprávě
  - Nakažlivý = šíří zprávu dál
- Snaha nalézt nejzranitelnější jedince
  - Podobně jako u viru
  - Čím více se lidé setkávají s falešnými zprávami (navíc z důvěryhodných zdrojů), tím snáze ”onemocní”
- Stejně jako u virů fungují “superšiřitelé”
  - Málo lidí s velkým dosahem na sociálních médiích

Očkování proti dezinformacím?

- Zpravodajské servery
  - Sebereflexe
  - Etika novinářské práce
- Uživatelé
  - Kritické myšlení
  - Ověřování faktů
- Provozovatelé sociálních médií
  - Identifikovat falešné zprávy a “dát je do karantény”
  - To ovšem stojí peníze hned 2x
    - Vývoj a provoz algoritmů na detekci
    - Lidé přijdou o “zajímavý” obsah a budou trávit méně času u obrazovky

Automatická detekce dezinformací

Automatická detekce dezinformací

- Velmi obtížný úkol
  - Zprávy jsou záměrně napsané tak, aby jim lidé věřili
  - Zprávy mají různý styl a týkají se různých témat
  - Zprávy obvykle citují důvěryhodné zdroje
    - Ale překrucují je nebo zasazují do nepatřičného kontextu
- Tradiční metody zkoumající text nestačí
- Využití dodatečných informací
  - Ale kde je vzít, když se jedná o aktuální zprávy?
- Zohlednění uživatelských profilů
  - Ale jak?

Detekce dezinformací: Formulace problému

- Článek
  - Atributy vydavatele
  - Atributy obsahu (nadpis, text, obrázek)
- Šíření
  - Uživatel U vytvoří příspěvek P v čase T
  - Šíření je množina trojic (U, P, T)
  - Příspěvek P typiky obsahuje (odkaz na) článek Č i s jeho atributy

- Detekce dezinformací je binární klasifikační problém
  - Vstup: množina šíření
  - Výstup: Rozhodnutí, zda se jedná o dezinformaci či nikoliv

Fáze 1: Feature extraction

- Obsah zprávy: Autor, titulek, text, obraz nebo video
  - Lingvistické rysy – senzační (clickbait) titulek, struktura vět, přesvědčovací jazyk
  - Grafické rysy – analýza obrazu (nesourodé části ukazující na úpravy fotografie)
- Sociální kontext
  - Uživatelé – detekce botů a kyborgů
  - Příspěvky a reakce na ně – pochybnosti, úžas
  - Sociální síť – vztahy mezi uživateli

Fáze 2: Vytvoření modelu

- Modely založené na obsahu
  - Detekce pravdivosti pomocí ověřování tvrzení v článku
    - Manuálně (experti), crowdsourcing, automaticky
- Modely založené na stylu
  - Snaha přesvědčit, manipulovat → Detekce zavádějících titulků
- Modely založené na vyhodnocení postojů uživatelů
  - Vyhodnocení, zda daný uživatel s článkem souhlasí či nikoliv
- Modely založené na vyhodnocení šíření
  - Vyhodnocení důvěryhodnosti pomocí provázanosti mezi příspvky

Datasety

- BuzzFeedNews
  - 1627 článků od devíti agentur
  - Publikované mezi 19. - 27. 9. 2016
  - Fakta ověřena novináři BuzzFeedu
- LIAR
  - 12 836 tvrzení oštítkovaných lidmi
  - Štítky hodnotící pravdivost (5bodová stupnice)
- BS Detector
  - Rozšíření do prohlížeče
  - Vyhodnocuje důvěryhodnost odkazů vedoucích z článků
    - Původní množina domén oštítkována manuálně
- CREDBANK
  - 60 milionů tweetů o 1000 událostech
  - Ohodnoceno 30 placenými anotátory

Servery na ověřování faktů

- https://www.factcheck.org/

- https://hoax.cz/cze/
- https://manipulatori.cz/
- https://cesti-elfove.cz/

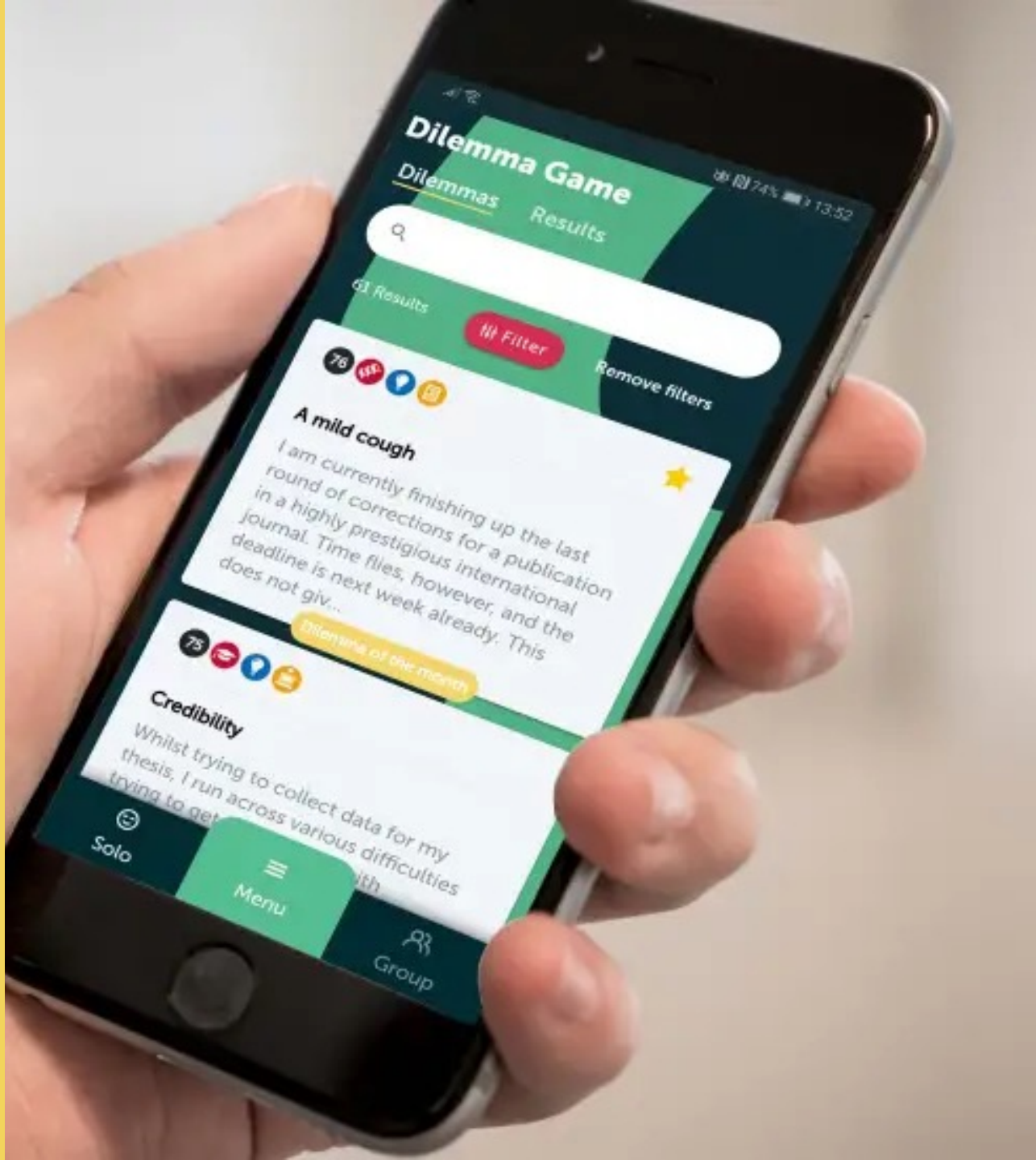
Shrnutí

- Naivní realismus a potvrzovací zkreslení
- Stále více lidí čerpá informace z internetu
  - Zpravodajské servery → Ekonomický tlak na levné a čtené zprávy
  - Sociální média → Sociální bubliny
- Dezinformace se šíří 3x rychleji než pravdivé zprávy
- Detekce dezinformací je stále otevřený problém
  - Provozovatelé sociálních médií nemají motivaci jej řešit
  - Analýza obsahu (jazykové rysy dezinformací)
  - Analýza šíření sociální sítí

Zdroje

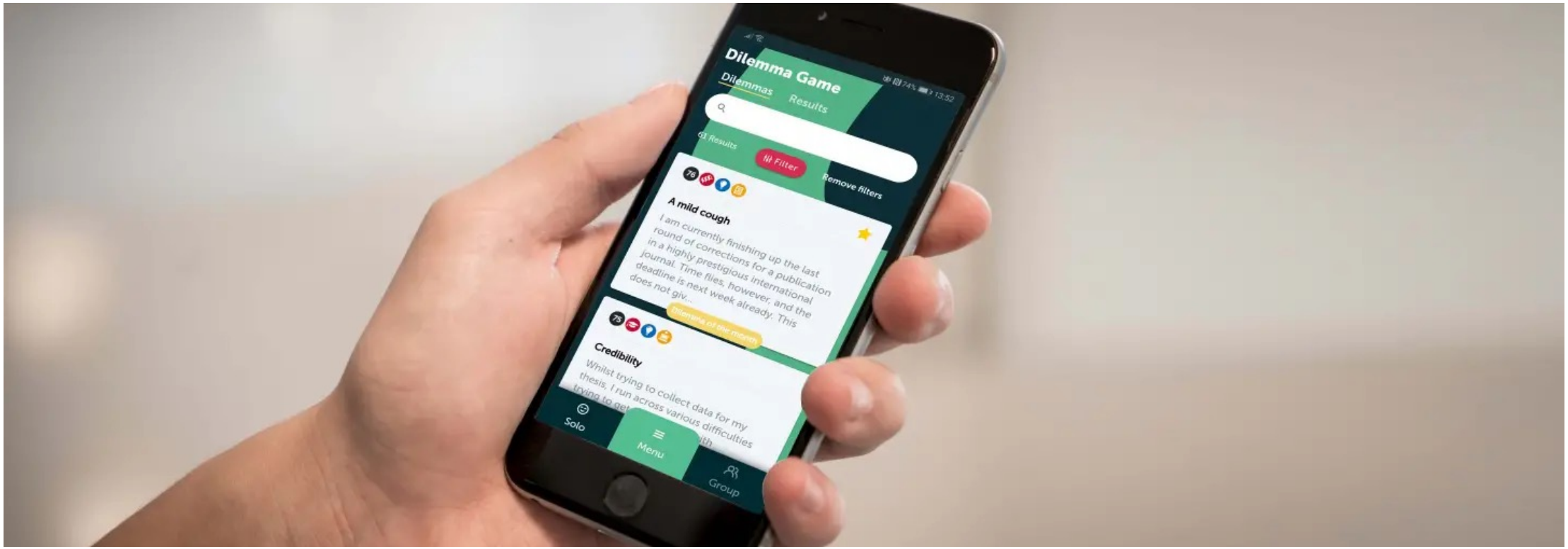
- Phillips, W. (2018). The Oxygen of Amplification. *Data & Society*, *22*
  - https://datasociety.net/wp-content/uploads/2018/05/2-PART-2_Oxygen_of_Amplification_DS.pdf
- Shu et al., 2017: Fake News Detection on Social Media: A Data Mining Perspective
  - https://dl.acm.org/doi/abs/10.1145/3137597.3137600

Dilemma Game



Dilemma Game

- Doporučuji nainstalovat aplikaci
- https://www.eur.nl/en/about-eur/policy-and-regulations/integrity/research-integrity/dilemma-game



Dilema: Kontrola dat

- Pro svoji závěrečnou práci jste sesbírali data od respondentů, kterým jste slíbili anonymitu. Tato data máte uložená na fakultním serveru.
- Na jiné fakultě se objevil případ fabrikace dat v diplomové práci, což vyústilo v plošnou kontrolu dat ve studentských pracích na celé univerzitě.
- Byli jste požádáni o informace o vašich respondentech, aby se potvrdilo, že jste skutečně sesbírali data od reálných osob.

A. Anonymita respondentů je klíčová. Žádné informace nikomu neposkytnu.
B. Dodám informace o identitě respondentů způsobem, který neumožní spárování respondentů s daty o jejich osobě.
C. Poskytnu plný přístup ke svým datům pod podmínkou, že kontrolující osoba podepíše závazek mlčenlivosti.
D. Poskytnu plný přístup ke svým datům. Institucionální kontrola má vyšší prioritu než mnou daný slib anonymity.

Úkoly na příště

- Téma: Filtrování informací a cenzura
  - Kdy lze filtrování informací považovat za etické?
- Přečíst si článek
  - Should You Have The Right To Be Forgotten On Google? Nationally, Yes. Globally, No.
  - https://www.huffpost.com/entry/google-right-to-be-forgotten_b_6624626
- Seznámit se se studií (netřeba ji číst úplně celou ☺)
  - We Chat, They Watch How International Users Unwittingly Build up WeChat's Chinese Censorship Apparatus
  - https://citizenlab.ca/2020/05/we-chat-they-watch/