Umělá inteligence jako zrcadlo společnosti
aneb jak zajistit, aby umělá inteligence rozhodovala spravedlivě?

,,Existuje skutečné nebezpečí pramenící ze systematizace diskriminace, kterou ve společnosti máme, prostřednictvím umělé inteligence. Myslím, že zatímco proplouváme světem plným neviditelných všudypřítomných algoritmů, musíme být velmi otevřeni a upozorňováni na míru chyb z této strany. " - *Timnit Gebru*

Citací vědkyně v oblasti umělé inteligence uvádím téma algoritmické korektnosti, objektivity a diskriminace s tím spojené. Představím dva způsoby špatné aplikace fascinujícího světa strojů i s jejich možnými řešeními.

Umělá inteligence umožňující zjednodušení každodenních aktivit jednotlivce i těžších úkolů společnosti se díky lidské důvěře šíří ohromně rychle. Stroj nemá emoce, řídí se pouze určenými příkazy a pravidly, není tedy ovlivňován subjektivními zájmy a prožitky. Automatickou reakcí člověka tudíž není aktivace kritického myšlení, naopak je mnohdy nahrazena předpokladem objektivity systému.
Je příjemné uplatňovat umělou inteligenci, kde to jen jde. Moderní společností je konstantně vylepšována, schopna nahradit člověka na mnoha místech, dnes dokonce i tam, kde je jí svěřena zodpovědnost samostatného rozhodování. K tomu musíme program *nakrmit* stovkami tisíc dat ve formě příkladů práce, kterou bude vykonávat, zahrnující vstupní i výslednou informaci. Neurony jsou poté schopny propojit jednotlivé situace a najít vzor pro vlastní rozhodování.

Představím dva způsoby chybné, nemorální nebo irelevantní, aplikace AI, díky kterým se ustálil pojem *computational bias:* systematický předsudek, naprogramovaný sklon umělé inteligence k privilegaci nebo znevýhodnění jednotlivců na základě sociální skupiny, do které spadají.
První problém pochází ze subjektivity vzorových dat, ze kterých se algoritmus učí. Jeho příkladem je přenechání prvotní selekce přihlášek na univerzitu nebo žádostí o práci počítačovým funkcím (Timnit Gebru, Salesforce). Algoritmus hledal systém v předešlých rozhodnutích, kde se mohlo objevit zaujetí například proti určité etnické skupině. Poté mohla být prvním kritériem jednotlivcova postupu barva pleti nebo gender. V takové situaci umělá

inteligence supluje pomyslné zrcadlo společnosti a vrozenou diskriminaci každého člověka zhmotňuje a systematizuje.

Další pochybení, které se může stát, je nevhodná aplikace statistik z minulosti k predikci budoucnosti. Vzniká ve chvíli, kdy člověk programu předá relevantní data, objektivní statistiky, které posléze funkce zohlední při rozhodování. Například studie (COMPAS, 2016) dokázala značné rozdíly mezi automatizovanou predikcí hrozby recidivy černochů a bělochů, v chybně pozitivní i chybně negativní. Vysvětlení, proč se to děje, je zcela jednoduché. Podíváme-li se na statistiky z roku 2008, příslušníci černošské komunity stálo za osmkrát více krádežemi než jiné rasy. Nikde tudíž nevznikla chyba ve zprostředkovaných datech.
Při posuzování člověka na základě sociální skupiny, do které se narodil, se jedná se o diskriminaci založenou na rase. Byť pravdivost teoretické predikce statistiky může být velmi pravděpodobná, nedává nám do ruky argument pro zadržení příslušníka dané skupiny, který mohl klidně spáchat stejné množství zločinů, jako příslušník bělošského etnika a být hodnocen přísněji.

Máme tu tudíž dva možné výskyty *computational bias.* První vzniká ve chvíli, kdy člověk sám promítne do programu své vlastní předsudky či zaujetí, nejedná se proto o relevantní data, ale o subjektivně zbarvená rozhodnutí. V druhém případě problém vykrystalizuje, když jsou relevantní data využita neetickým způsobem. Rozhodování podle příslušnosti k sociální skupině je nemorální a porušující zákon zavazující společnost k nezohledňování rasy, náboženství, či genderu při zacházení s druhými. Oba dva případy jsou v rozporu s antidiskriminačním zákonem.

Přítomnost *computational bias* je ale celospolečensky nepříznivá i při pohledu do budoucnosti. Postupem času, a to zejména v posledních desítkách až stovce let, se společnost zasazuje o rasovou i genderovou rovnost. Snažíme se udělat vše proto, aby si byli všichni právně rovni, ale i pro limitaci a maximální eliminaci předsudků zakořeněných v našem světě. Pokud se vrátím k příkladu černošské kriminality, mohu předvést, že některé předsudky mají pochopitelný základ.
Systém, ve kterém využíváme statistik z minulosti k předsudkům do budoucnosti, nás vtahuje do začarovaného kruhu. Množství kontrol roste spolu s množstvím policejních záznamů, zároveň kvantita policejních záznamů je logicky vyšší u skupiny, která je kontrolována častěji. Pokud necháme umělou inteligenci podporovat tyto rozdíly, upozorňovat na ně,

logicky se bude zvyšovat rozdíl v etnickém výčtu zadržených a diference mezi sociálními skupinami se bude pouze prohlubovat a hlavně systematizovat.

Teď k té nejdůležitější otázce: jak můžeme takový problém řešit? Naším cílem je, aby naprogramované funkce při rozhodnutích o budoucnosti občanů nebyly ovlivněny sociální skupinou, do které se narodili.

Pro pochopení řešení se musíme ponořit hlouběji do fungování umělé inteligence. Základem AI systémů jsou zpravidla neuronové sítě, které jsou buďto kompletně transparentní, člověk se může podívat na přesný proces, kterým prošla původní informace před přeměnou do výchozí. Opakem je síť, která funguje na principu *black box,* známe tedy pouze původní informaci a konečný produkt.
Transparentní programy, tzv. opensource, jsou počítačoví pracovníci schopni rozložit na malé části, tudíž je možné najít přesnou oblast, kde systém dochází k diskriminačním závěrům. U uzavřených neuronových sití může člověk kontrolovat její rozhodnutí pouze skrze původní informaci (byť už jsou v procesu výzkumy, jejichž cílem je i do takových proniknout[1]).

Nezbytná je proto kompletní přístupnost dat, která jsou předávána algoritmům, komukoli. Pokud chceme, aby každý plnil svou funkci ve snaze eliminace diskriminace, musí na tom najít svůj zájem i ti, kterých se problém netýká. Pokud by měl kdokoli možnost nahlédnout do systematičnosti oblastí, ve kterých se algoritmické zaujetí vyskytuje, bylo by jednodušší odhalit diskriminační postupy společnosti, protizákonné chování, nebo pouze chybnou aplikaci (např. nepřesnost demografických skupin).

Computational bias začíná pronikat do čím dál více společenských oblastí, které mají moc ovlivňovat lidské osudy. I to na ně upoutává pozornost nejen skupin, kterých se problém týká. Algoritmické zaujetí vzniká buďto jako následek lidského pochybení nebo ve chvíli, kdy člověk nechá svůj sklon k zaujatosti ovlivnit vkladové informace, případně nemorálním vytvářením předpokladů na základě statistik. Obě formy mohou jít proti antidiskriminačnímu zákonu a pouze prohlubují a systematizují propasti a rozdíly mezi společenskými skupinami.
Kromě upozorňování na problém by měly být funkce transparentní a přístupné pro pravidelné kontroly, mělo by se apelovat na eliminaci algoritmického zaujetí. V praxi potom můžeme ovlivňovat jednotlivé části sítě, kde se *bias* vytvořil, ale u některých zařízení máme moc pouze nad daty, která vkládáme, proto je kontrola nezbytná.

[1]https://towardsdatascience.com/explainable-artificial-intelligence-14944563cc79
https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/?sh=23ab4e87c9ef
https://www.salesforce.com/video/3402966/
https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/
https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/
https://www.vox.com/science-and-health/2019/1/23/18194717/alexandria-ocasio-cortez-ai-bias
https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen
https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=1691&context=facpub
https://www.youtube.com/watch?v=N9XaLNfExgM
https://www.youtube.com/watch?v=p-82YeUPQh0
https://news.un.org/en/story/2020/12/1080192
https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses
https://www.nature.com/articles/d41586-019-03228-6
https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73