**Jak zajistit, aby umělá inteligence rozhodovala spravedlivě?**

Umělá inteligence představuje jeden z nejvíce užívaných a rozvíjených počítačových oborů posledních let. Bezpochyby umělá inteligence přináší spoustu dobrého do našich životů. Od pokroku ve zdravotnictví, přes vytváření přesnějších ekonomických modelů až po pomoc s bojem proti světovému hladu[1]. S možnostmi, které si ještě ani nedokážeme představit, nám dává velkou moc, ale zároveň přináší i rizika. Může ale umělá inteligence (ne)úmyslně některé skupiny osob diskriminovat a tím porušovat jejich lidská práva?

Umělá inteligence neboli AI funguje na mnoha principech (neuronové sítě, strojové učení, evoluční algoritmy atd.), a i přestože se v mnoha věcech tyto druhy AI liší, mají společnou jednu věc, a to jsou data, ze kterých se učí. Na základě těchto dat je umělá inteligence schopna hledat korelace, odvozovat vzorce, skoro až předpovídat budoucnost, dělat za nás rozhodnutí a zejména se učit a sama sebe rozvíjet.[2]

Problém nastává, když umělá inteligence nefunguje stejně pro všechny a v nějakém sektoru má tendence určitou skupinu osob upřednostňovat. Například v bankovnictví má AI tendence posilovat sociální nerovnost a dávat horší úroky jen na základě rasy.[3] V USA byl Afroameričan vězněn na základě špatného provedení rozpoznání obličeje policejním softwarem.[4] Software pomáhající v soudnictví určovat recidivisty jednal také zaujatě k černochům. Program na popisování fotek označil společnou fotografii jedné rodiny jako „skupinu šimpanzů“.[5] V Británii zase studenti neměli stejné podmínky pro přijetí na univerzitu, protože počítačový algoritmus znevýhodňoval studenty z chudších čtvrtí. Studie z roku 2016 u softwaru na rozpoznávání hlasu od Googlu (v té době nejlepšího na trhu) zjistila, že je o 70% větší pravděpodobnost, že software přesně rozezná mužskou řeč než ženskou. To může ženy značně ohrozit na zdraví, jelikož tyto programy na rozpoznávání řeči jsou využívány v bezpečnostních systémech některých aut a ve zdravotnictví. V jiných případech zase software hodnotící CV

[1] Umělá inteligence je využívána například k maximalizaci produkce plodin a k dřívější detekci nemocí.

[2] European parliament. *What is artificial intelligence and how is it used?* [online]. 2020-10-22 [cit. 2020102-13]. Dostupné z: https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

[3] Council of Europe. *Discrimination, artificial intelligence, and algorithmic decision-making* [online]. Strasbourg: Directorate General of Democracy, 2018. 36 s. [cit. 2021-02-13]. Dostupné z: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

[4] UN News. *Bias, racism and lies: facing up to the unwanted consequences of AI* [online]. 2020-11-30 [cit. 2021-02-13]. Dostupné z: https://news.un.org/en/story/2020/12/1080192

[5] ŠTĚDROŇ, Bohumír a kol. *Právo a umělá inteligence*. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, s.r.o., 2020. 16 s.

nedával ženám stejné příležitosti, přestože byly na danou práci kvalifikovanější.[6] Jsou zde ale opravdu v ohrožení naše lidská práva?

Všeobecná deklarace lidských práv uvádí, že všichni lidé se rodí sobě rovní, co do důstojnosti a práv bez jakéhokoli rozlišování, zejména podle rasy, barvy, pohlaví, jazyka a náboženství.[7] Jakmile je někdo znevýhodněn u soudu na základě jeho rasy, rázem už neplatí ani článek 7 a to, že všichni jsou si před zákonem rovni, bez jakéhokoli rozlišování. Deklarace nepovoluje diskriminaci ve vztahu k lidským právům, existují ale i jiné dokumenty, které diskriminaci zakazují (v ČR je to antidiskriminační zákon[8], v USA Civil Rights Act of 1964[9]). Půjčku nemusí dostat každý, ale tím, že je člověku upřena půjčka pouze na základě rasy, jedná se o protizákonné jednání, které nesmíme tolerovat od lidí ani od počítačů.

Zároveň je zde ale důležité podotknout, že umělá inteligence není sama o sobě rasistická, pouze vykazuje v některých oblastech velkou chybovost nebo nepřesnost pro určitou skupinu lidí. Nikdo úmyslně neprogramuje algoritmy tak, aby měly menší spolehlivost u žen či Afroameričanů. Snahou každé umělé inteligence je, aby měla co největší úspěšnost, což v některých případech pro některá vstupní data se nemusí podařit.

Hlavním důvodem, proč tato chybovost vzniká je tzv. „Data Gap“, neboli chybějící data.[10] Jak jsem již zmínila, umělá inteligence vychází z milionů vzorků dat, na kterých se učí „přemýšlet“. S tím ale přichází problém, protože vzorky dat ne vždy mohou pokrýt celou cílovou skupinu. Jakmile se AI učí pouze na určitém vzorku, nejčastěji na bílých mužích, a je poté testována na vzorku dalších bílých mužů, jak pak může spolehlivě fungovat pro ženy nebo Afroameričany? Nemůže. Samozřejmě dává smysl, že pokud budu vyvíjet program určený pro ženy, nezahrnu do zkoumaného vzorku muže. Pokud ale se jedná o něco, co v budoucnu má být využíváno a má fungovat napříč všemi obyvateli, musí se do vstupních dat zahrnout všichni. O to více to platí na programy, na kterých závisí lidské životy, jako jsou například ty užívané ve zdravotnictví a soudnictví. Jak bychom tedy měli naložit s chybovými rozhodnutími umělé inteligence, které často může působit jako zaujaté?

Jak bylo řečeno, chybovost většinou vyplývá z nedostatečně různorodého vzorku vstupních dat a tím pádem AI nemůže přesně předpovídat. Tento problém by se dal vyřešit tím, že pokud by

---

[6] PEREZ, Caroline Criado. *Invisible Women: Data Bias in A World Designed for Men*. Londýn: Chatto&Windus, 2019. 162, 166-167 s. ISBN 978-1-78474-172-3.

[7] Všeobecná deklarace lidských práv, OSN, 1948. Článek 1 a 2.

[8] Zákon č. 198/2009 Sb., o rovném zacházení a o právních prostředcích ochrany před diskriminací a o změně některých zákonů (antidiskriminační zákon). In: *Sbírka zákonů České republiky,* 2009.

[9] Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000e *et seq* (1964).

[10] PEREZ, Caroline Criado. *Invisible Women: Data Bias in A World Designed for Men*. Londýn: Chatto&Windus, 2019. XII s. ISBN 978-1-78474-172-3.

firma vyvíjela software „pro všechny“, musela by uvést z jakého vzorku se AI učí a jak je program efektivní pro různé skupiny obyvatel. Poté by už bylo na zodpovědnosti každé instituce, která program rozhodne využívat, aby zvážila limity daného softwaru a pro jakou skupinu obyvatel ho bude využívat. Například pokud policejní systém na rozpoznávání obličejů bude mít velkou chybovost u lidí tmavé pleti, je na policii, aby rozvážně využila jeho možností, ale zároveň je nutné, aby i vývojáři systému dostatečně informovali uživatele.

Co je ale zásadní, je samotný přístup při programování AI. Měl by být kladen velký důraz na to, aby data byla sbírána i z nejchudších oblastí světa, a byli v nich zahrnuti lidé napříč všemi vrstvami obyvatelstva i regionů, a tudíž, aby celý systém nebyl naprogramován a vyladěn na průměrného bílého muže. Zároveň toto ale může být v některých případech nerealizovatelné anebo finančně příliš náročné a firmám se to ne vždy vyplatí.

Myslím, že jako státní instituce kladou velký důraz na testování léčiv a očkování a jejich bezpečnost a účinnost, tak stejným způsobem by měli testovat software, který je využíván ve veřejném sektoru. Tato kontrola by odhalila slabiny programů a poté by umožnila se zjištěnou chybovostí dále pracovat a nastavit nějaké korekční procesy.

Další věc, která by mohla pomoci problematice umělé inteligence a její spravedlnosti je větší dohled nad rozhodováním umělé inteligence a možností se odvolat. Nyní už sice v některých státech existuje vyhláška, že pokud nějaké rozhodnutí bylo vytvořeno čistě na základě AI, daný člověk o tom musí vědět a může se odvolat, ale v praxi to zatím moc nefunguje a oběti zatím nemají dostatečnou moc proti systému bojovat.[11]AI dělá také chyby, ale problém je v tom, že u umělé inteligence, na rozdíl od lidí, je těžké určit, kdo nese zodpovědnost za její činy a je těžké najít viníka. O to důležitější by měl být dohled nad AI, což se spíše nedaří, protože se moderní technologie rozvíjí tak rychle, že je nestíháme právně regulovat.

Lidská práva a obecně zákony se nesmí porušovat, ať už se jedná o pochybení ze stran lidí nebo počítačů. Musíme se proto snažit co nejrychleji zaplnit „Data Gap“, aby se už nestávalo, že umělá inteligence bude fungovat pouze pro jednu část z nás a tím znevýhodní zbytek obyvatel. Jelikož AI na základě dat zkoumá chování naší společnosti, která často (ne)úmyslně smýšlí na základě historicky zakořeněných genderových a rasových předsudků, měli bychom si k srdci opravdu vzít kategorický imperativ Immanuela Kanta: „*Jednej podle zásady, o které bys chtěl,*

[11] BURANYI, Stephen. *Rise of the racist robots – how AI is learning all our worst impulses*. In: The Guardian [online], 2017-08-08. [cit. 2021-02-13]. Dostupné z: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

*aby se stala obecným zákonem.*“[12] a uvědomit si, že nyní se naše chování opravdu stává obecnými zákony, podle kterých se řídí celý svět technologií.

[12] KANT, Immanuel. *Základy metafyziky mravů*. Praha: 1990, str. 83.