Základné etické výzvy a riziká systémov umelej inteligencie v optike vzdelávania[1]

ThLic. Ing. Peter Šantavý, PhD.

Slovo na úvod (slide 1)

Fenomén umelej inteligencie patrí k technologickým celebritám súčasného sveta. Systémy umelej inteligencie pravdepodobne už prekročili pomyselnú hranicu praktickej využiteľnosti, množstva oblastí nasadenia a predikovanej ekonomickej návratnosti i očakávaných benefitov. Nástup týchto technológií sa preto javí ako nevratný a v rámci znalostnej spoločnosti v blízkej budúcnosti prakticky všadeprítomný.

Už zmeny, ktoré prináša informatizácia a digitalizácia so všetkou virtualizáciou, kybernetickým priestorom a podobne, môžeme nazvať paradigmatickými. To znamená, že informačné technológie a dáta sa z roviny prostriedku dostávajú do roviny kontextu spoločnosti až do tej miery, že na základe zmien v technológiách, vzhľadom na prevratný vedecký i technologický pokrok, sa mení medziľudská komunikácia v esenciálnej rovine. Teda menia sa vzťahy, mení sa spoločnosť a princípy jej fungovania.

Fenomén umelej inteligencie tieto zmeny nielen prehlbuje a akceleruje, ale prináša i viaceré nové výzvy. Napríklad stroj sa stáva subjektom komunikácie, vplyv technológií má dopad na kognitívne schopnosti a emocionálnu stránku človeka, mení sa vzťah medzi inteligenciou a racionalitou a podobne.

Naviac sa to všetko deje v rámci postmodernej rezignácie na rácio a vo víre hodnotovej relativizácie.

Umelá inteligencia ako disruptívna a emergentná technológia[2] má potenciál ovplyvňovať vývoj človeka, extrémne vplývať na jeho psychiku, zmýšľanie a činnosť, zasahovať do fungovania spoločnosti a atakovať viaceré oblasti ľudských práv. Preto fenomén umelej inteligencie bytostne presahuje technologickú stránku veci.

---

1 Text primárne čerpá z 1. vydania mojej publikácie Umelá inteligencia – dobrý sluha a zlý pán? a z jej pripravovaného 2. vydania.
ŠANTAVÝ, P. *Umelá inteligencia – dobrý sluha a zlý pán?* [on-line]. Bratislava: RKCMBF UK, 2023. ISBN 978-80-88696-91-9.
Dostupné na internete: <https://peter.santavy.cloud/ai-good-servant-and-bad-master>

2 Disruptívna a mergentná – nová a rozvíjajúca sa technológia, ktorá nahrádza a ruší technológie predošlé a v mnohom mení spôsob, akým spoločnosť funguje.

Čím komplexnejšie a sofistikovanejšie systémy umelej inteligencie máme a čím väčší dopad na život človeka a spoločnosti si uvedomujeme, resp. už zakúšame, tým viac sa formuje celosvetové úsilie o ich etické a bezpečné využívanie.

Výziev – osobitne tých etických – spojených s nástupom umelej inteligencie je neúrekom.

1

Základné etické výzvy a princípy systémov AI (slide 2)

V pohľade na niekoľko desaťročí rozvoja oblasti moderných informačných technológií si uvedomujeme, že prakticky ešte nikdy nebolo badať toľké obavy z nasadenia novej technológie a tak seriózny záujem o etické otázky ako v prípade systémov umelej inteligencie. Kľúčové slovo či značka (hashtag) #AIEthics v komunite odborníkov neoznačuje okrajovú záležitosť, ktorá je mimo zorného uhľa pohľadu tých, čo umelej inteligencii skutočne rozumejú, ale stáva sa súčasťou hlavného prúdu vývoja, implementácie a používania systémov AI v rámci reálneho sveta.

Aby sme si rozumeli, s futurologickými obavami takmer na úrovni sci-fi sa spoločnosť stretáva už od pionierskych čias formovania konceptu umelej inteligencie a prvých krokov na poli vývoja týchto systémov. V našom prípade však ide o seriózny záujem celej spoločnosti, ktorá je konfrontovaná s technológiami, ktoré majú potenciál hlboko zasahovať a ovplyvňovať jednotlivcov i celé spoločenské celky.

Tento seriózny záujem sa neobmedzuje len na oblasť umelej inteligencie, ale má oveľa širší záber, ktorý koreluje a má i kauzálnu súvislosť už s treťou priemyselnou revolúciou, digitálnym vekom a zmenou paradigmy nastupujúcej informačnej spoločnosti, v rámci ktorej sa na základe zmien v technológiách a vzhľadom na prevratný vedecký i technologický pokrok mení medziľudská komunikácia v esenciálnej rovine, menia sa vzťahy, mení sa spoločnosť a princípy jej fungovania.

V intenciách narastajúceho záujmu celej spoločnosti o problémy, ktoré ju bytostne ovplyvňujú, je aj osobitné akcentovanie etického rozmeru vývoja, tvorby, nasadenia, poskytovania a využívania celého spektra technológií umelej inteligencie.

Toxický potenciál technológií AI (slide 3)

Existujú len dva druhy priemyslu, ktoré svojim zákazníkom hovoria používatelia. Nelegálne drogy a softvér. (Prof. Edward R. Tufte)

Jeden z citátov, použitých vo filmovom dokumente The Social Dilemma[3] o rizikách sociálnych sietí a za nimi stojacich algoritmov umelej inteligencie, nás provokuje si **uvedomiť možný toxický potenciál technológií skrytých za fenoménom umelej inteligencie a potrebu adekvátneho riešenia celého cyklu činnosti týchto systémov** – teda ich etického vývoja, výroby, nasadenia, prevádzky, využívania a bezpečného ukončenia činnosti.

Na dobro človeka zameraná umelá inteligencia (slide 4)

Základným princípom pre akýkoľvek systém umelej inteligencie je zameranie na dobro človeka, teda známy a všeobecne prijímaný princíp human-centered and beneficial artificial intelligence.

Je to princíp, o ktorom síce takmer všetci v oblasti vývoja, nasadenia a využívania týchto technológií hovoria, no **mnohokrát tým niečo diametrálne odlišné chápu**.

Na dobro človeka zameraná umelá inteligencia konkrétne (slide 5)

Treba zdôrazniť, že princíp umelej inteligencie zameranej na človeka by mal:

- byť chápaný v duchu **klasickej filozofickej antropológie** (predovšetkým biologickej a kultúrnej);
- zahŕňať **každú** ľudskú bytosť a **nikoho** nediskriminovať;
- mať na zreteli dobro **ľudstva** a spoločnosti, chrániac pri tom a rešpektujúc dobro každej ľudskej bytosti;
- sa vyznačovať starostlivosťou o náš „spoločný a zdieľaný domov“, teda o **celý svet**.

Chápeme, že human-centered prístup zahŕňa aj mnohé technologické, praktické a právne náležitosti vývoja a nasadenia systémov umelej inteligencie, no bez vyššie uvedeného návrhu aplikácia tohto princípu môže podliehať relativizmu a erózii hodnôt, či postupnému vyprázdneniu jeho podstatných aspektov.

---

3 ORLOWSKI, J. *The Social Dilemma*. [filmový dokument]. Netflix, 2020. [cit. 7. decembra 2021]. Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

Rámec klasickej filozofickej antropológie a z neho prameniaci diapazón humanizmu nám poskytuje aj dôležité morálne vyhranenie sa voči pokušeniu umelo vylepšovaného človeka prostredníctvom umelej inteligencie, čím by sme otvorili Pandorinu skrinku relativizácie hodnoty každej ľudskej bytosti, eugeniky, transhumanizmu a pod.

Interdisciplinárny rámec ako základ (slide 6)

Skutočné riešenie etických problémov a výziev technológií umelej inteligencie nie je možné úspešne realizovať bez interdisciplinárneho rámca, v rámci ktorého sme dostatočne oboznámení aj s technologickou stránkou týchto systémov a psychologickými, sociologickými i právnymi aspektmi ich nasadenia.

Dostatočné oboznámenie sa **s technologickou stránkou** nám predovšetkým umožňuje pochopiť podstatu a rozsah technologických limitov a rizík algoritmov AI, a následne si v spolupráci s ďalšími odbormi adekvátnejšie predstaviť ich dosah na jednotlivé oblasti reálneho nasadenia, dôsledky na život človeka a fungovanie spoločnosti.

Identifikácia oblastí, v ktorých sa seriózne nasadenie systémov umelej inteligencie nezaobíde bez uspokojivého návrhu riešenia etických problémov, tiež vyžaduje dôkladnú **analýzu psychologických, sociologických i právnych aspektov ich nasadenia**.

Rozsah diskutovaných tém môže byť skutočne rozsiahly – napr. sociálna spravodlivosť, ľudské práva, spravodlivá vojna, bioetické otázky, transhumanizmus,...

V horizonte vývoja všeobecnej umelej inteligencie môžu mať nezastupiteľnú úlohu viaceré **netechnické vedecké disciplíny**, schopné prispieť k riešeniu modelov správania, teórie mysle, simulácie emócií i vedomia a pod.

K dôležitým rozmerom interdisciplinárnej komunikácie patrí aj **spolupráca s inštitúciami na medzinárodnej i štátnej úrovni**, bez ktorej nie je možné vytvárať reálne etické a právne regulačné rámce, a v neposlednom rade kooperácia s privátnym sektorom, ktorého vývojové kapacity a finančné i ľudské zdroje sú motorom vývoja a nasadenia systémov AI takmer do všetkých oblastí reálneho života.

Osobitným aspektom interdisciplinárneho rámca je vzdelávanie, osveta, prevencia a formácia, čo však ešte spomenieme v ďalšom texte.

Dôveryhodná umelá inteligenca (slide 7)

Základným kritériom pre principiálne etické nastavenie návrhu, vývoja, implementácie, nasadenia a využívania týchto technológií AI je princíp orientácie na dobro človeka, ktorý sme už spomínali.

V reále sa principiálny postoj orientácie na dobro človeka stáva ekvivalentným problematike dôveryhodnosti umelej inteligencie, pričom je treba stanoviť podmienky, bez splnenia ktorých by nasadenie systémov do reálneho sveta, v ktorom pôsobia na človeka a vplývajú na spoločnosť, nemalo byť umožnené.

Dôveryhodné systémy musia byť **legálne**, teda musia vyhovovať požadovaným normám a spĺňať všetky platné zákony, predpisy i regulácie. Musia byť **etické**, teda musia rešpektovať etické zásady a hodnoty. A musia byť **robustné**, čo znamená, že musia dosahovať potrebné štandardy bezpečnosti a spoľahlivosti nielen z technologického hľadiska, ale zohľadňovať aj sociálne prostredie a dopady na spoločnosť.

Dôveryhodná umelá inteligencia – praktické dôsledky (slide 8)

Praktickým dôsledkom princípu dôveryhodnosti je stanovený základný etický rámec a zákonné regulácie pre tvorcov, poskytovateľov a používateľov týchto systémov i implementované eticko-právne požiadavky a obmedzenia priamo v týchto systémoch.

(Nielen) angažovanosť, legislatívne kroky a regulácie (slide 9)

Súčasné iniciatívy, ktoré prebiehajú vo vládnych a vedeckých pracovných skupinách i technologických a komunitách, sa snažia formulovať nielen potenciálne prínosy umelej inteligencie, ale súčasne i identifikovať a obmedzovať jej riziká.

Až donedávna v žiadnej z krajín sveta neexistovala ucelená legislatíva pokrývajúca celú problematiku umelej inteligencie. Doterajšie regulácie dotýkajúce sa technológií AI boli buď veľmi špecifické, riešiac parciálne problémy konkrétnych oblastí nasadenia týchto systémov a / alebo boli súčasťou iných regulácií, napr. kybernetickej bezpečnosti, ochrany osobných údajov, sektorových regulácií v rámci finančného sektora alebo štátnej správy.

V rámci Európskej únie však vzniká úplne nová regulácia – **Nariadenie o umelej inteligencii**, ktorý nemá obdobu nikde vo svete. Stanovujúc si za cieľ pokryť takmer celú problematiku súčasných systémov umelej inteligencie, ide vôbec o prvú komplexnú reguláciu konkrétnej technológie.

Nariadenie o umelej inteligencii sa tak stáva jedným z vrcholov snáh o riešenie dôveryhodnej AI. Samozrejme, nie je to jediný počin – v prezentácii uvádzame aspoň náznakovo niekoľko súčasných aktivít v oblasti regulácie a etiky.

Žiaľ, vzhľadom na už existujúci ekonomickú návratnosť, pomerne dobre fungujúce biznis modely i očakávané benefity **badať aj opačný trend – snahu obmedziť regulačné**

aktivity a etické záväzky vo vidine ekonomického zisku alebo technologickej nadvlády s jej dôsledkami.

Príkladom môže byť nedávna tzv. *Noc dlhých nožov* v OpenAI alebo licitácia podmienok a výsledného znenia Nariadenia o umelej inteligencii na pôde EÚ i lobovanie veľkých technologických korporácií proti tejto regulácii (napr. CCIA).

Žiaľ, v súčasnosti nemôžeme očakávať prioritné zameranie sa technologických spoločností na dodržiavanie etických princípov. Okrem vojenskej oblasti a oblasti riadenia štátu, ktoré majú svoje vlastné priority, je základným cieľom orientácia na finančný zisk.

Príklad štruktúry etických požiadaviek a oblastí osobitného záujmu (slide 10/11)

Uvádzame konkrétny príklad štruktúry etických požiadaviek, možností angažovania sa a oblastí, ktorým sa treba osobitným spôsobom venovať.

Dôveryhodná umelá inteligenca a ľudský faktor (slide 12)

Keďže dopad na spoločnosť bude nesmierny, jednou z priorít sa javí nielen potrebné vzdelávanie, osveta a formácia morálnych postojov vývojárov, poskytovateľov i používateľov technológií umelej inteligencie, ale i edukácia a osveta celej spoločnosti, bez ktorej si ťažko predstaviť rast spoločenskej citlivosti a zodpovednosti v oblasti celoplošnej adaptácie a využívania týchto systémov.

2

Riziká a limity súčasných systémov AI (slide 13)

Riziká a limity súčasných systémov sa podieľajú na mnohorakých problémoch, ktoré so sebou technológie umelej inteligencie prinášajú – či už ide o zlyhania technického rázu alebo dôsledky atakujúce ľudskú psychiku, hodnoty, ľudské práva a celú spoločnosť.

Fiktívny rozhovor s ChatGPT (slide 14)

Všetci, ktorým sci-fi sága Terminátor niečo hovorí, chápu, o čo ChatGPT v tomto fiktívnom rozhovore ide. **Aj keď si uvedomujeme, že ide o vtip, nepríjemná pachuť neistoty a nedôvery zostáva...**

Technologické rizikové faktory (slide 15-21)

Pri vývoji, realizácii, nasadzovaní a využívaní technológií AI je treba rátať s viacerými obmedzeniami a rizikami súčasných systémov umelej inteligencie.
Nasledovné slajdy (15-21) predstavujú výber dôležitých technologických rizík a problémov, ktoré môžeme zhrnúť v zásade do piatich oblastí:

- datasety
- modely a hyperparametre
- technologické a procesné riziká a limity
- vysvetliteľnosť a interpretovateľnosť (XAI)
- nasadenie a využívanie

Vybrané dôsledky a riziká pre človeka (slide 22-27)

V modernej informačnej spoločnosti sa základnou komoditou stávajú informácie. Systémy umelej inteligencie sú priamo závislé na kvalitných informáciách, pričom o väčšine aktuálne najúspešnejších sofistikovaných systémoch AI sa dá povedať, že sú závislé na extrémnom množstve relevantných dát.

Aby nasadenie systémov umelej inteligencie bolo v spoločnosti úspešné, potrebné dáta musia byť neustále zhromažďované z reálneho sveta a priamo z ľudského prostredia.

Nutnou súčasťou extrémne rýchleho súčasného vývoja v oblasti umelej inteligencie je fungujúca možnosť monetizácie nasadenia systémov AI, čo sa najmarkantnejšie prejavuje v on-line systémoch a produktoch technologických gigantov (sociálne siete, vyhľadávače,...). U týchto systémov by sa pri povrchnom pohľade mohlo javiť, že základnou komoditou sú informácie, ktoré uvedené spoločnosti bezprecedentným spôsobom zhromažďujú, analyzujú a spracúvajú. Avšak systémy umelej inteligencie so svojimi schopnosťami posúvajú túto časť paradigmatickej zmeny informačnej spoločnosti ešte ďalej: **z informácií a z produktov ich spracovania sa komoditou stávajú priamo ľudia – ba čo viac, ich psychologické profily a vzťahy, konanie a jeho ovplyvňovanie a v konečnom dôsledku i fungovanie celej spoločnosti**.

~~Dokonale~~ (povedzme, že len tak kvalitne, ako to umožňujú viaceré súčasné platformy sociálnych sietí) natrénovaná AI sa dokáže zamerať na človeka, konkrétne sociálne skupiny, resp. vo všeobecnosti nás ľudí v spoločnosti a – analogicky využitiu kvalitných psychologických metód – priviesť nás k pozornosti, akceptácii ňou predkladaných informácií (osobitne cez pokročilé metódy, napr. augment reality, podprahové metódy,...) a k **ovplyvňovaniu nášho vnímania, našich rozhodnutí a konania**... A môže to ísť až tak ďaleko, že prichádza k zmene nášho zmýšľania a vnímania, kým vlastne sme. Druhou stranou mince je **schopnosť systémov AI vytvárať modely, ktoré sú schopné predpovedať naše konanie**.

Ak efekt má byť takmer istý, treba stavať na extrémne presných predikciách. A extrémne presné predikcie nie je možné vykonávať bez extrémne veľkého množstva dát1 a techník, ktoré ich dokážu v reálnom čase spracovať.

S tým súvisí ďalšie riziko – tzv. kapitalizmus dohľadu (surveillance capitalism), ktorý ťaží z nekonečného sledovania – **ľudia i celá spoločnosť sú vystavení neustálemu dohľadu a sledovaniu, ktoré je však len veľmi málo pod kontrolou, ak vôbec**. Osobitne tak môže byť v kontexte systémov umelej inteligencie, u ktorých zhromažďované dáta nie sú takmer vôbec pod ľudským dohľadom (ak taký dohľad je pri veľkých systémoch zhromažďujúcich denno denne extrémne množstvo dát2 vôbec možné zaručiť).

Ide o začarovaný kruh, čím viac dát systémy AI dokážu získať (a na nich sa učiť), tým presnejšie modely nášho správania ponúkajú a tým lepšie dokážu ovplyvňovať naše vnímanie, postoje a rozhodnutia – rozhodnutia i činnosť, čo následne generuje ďalšie dáta, zberané a využívané systémami AI... Na základe dát a výsledkov systémov AI sa upresňujú algoritmy implementované v týchto systémoch, aby požadované výsledky boli stále presnejšie (a modely nášho správania autentickejšie).

Neuvedomujúc si, ako je naša myseľ a psychika zraniteľná, tak postupne prechádzame od technologického prostredia založeného na systémoch AI k prostrediu založenému na závislosti a manipulácii.

Potenciál a skutočnú silu týchto psychologických zásahov, vyjadruje i skúsenosť ľudí pracujúcich v oblasti implementácie systémov AI v informačnej spoločnosti, osobitne v rámci sociálnych sietí. Mnohí z nich do detailov poznajú spôsob fungovania nasadených systémov AI, poniektorí z nich dokonca tieto systémy vyvíjali a nasadzovali, no sami sa stali obeťami ich fungovania. Otvorene hovoria o obdobiach svojho života, v ktorých zakúsili totálnu neschopnosť vymaniť sa z ich vplyvu a závislosti.

Dr. Anna Lembke, lekárska riaditeľka liečby závislostí na Stanfordskej univerzite, poukazuje na potenciál závislosti, ktorú môžu vytvárať moderné sociálne siete, postavené na systémoch AI a dostatočne sofistikované, aby virtualizovali a nahrádzali naše reálne vzťahy: „Sociálne médiá sú droga. Máme základnú biologickú potrebu byť v kontakte s inými ľuďmi, čo má priamy vplyv na uvoľňovanie dopamínu v mezilimbickej dráhe. Tento systém, ktorý majú na svedomí milióny rokov evolúcie, sa nás snaží spájať a núti nás žiť v komunitách, aby sme si mohli nájsť partnerov a množiť sa. Takže niet pochýb o tom, že niečo ako sociálne médiá, ktoré umožňujú spojenie medzi ľuďmi, budú mať potenciál vyvolať závislosť.“

So závislosťou prichádzajú aj ďalšie súvisiace problémy. Algoritmy sociálnych médií sú postavené na neustálych podnetoch, ktoré súvisia s uvoľňovaním dopamínu a pozitívnou reakciou našej mysle. Vytvára sa tak túžba po neustálom pozitívnom ohodnotení, ktoré už nie je primeranou a hlavne adekvátnou súčasťou vnemov v reálnom svete, ale umelo držanou hladinou, na ktorej sa používatelia stávajú závislí. Znižuje sa tak schopnosť prijímať a adekvátne spracovávať negatívne reakcie a prijať aj svoje slabé stránky. A v konfrontácii s realitou alebo s negatívnou skúsenosťou prichádza k psychickým problémom a kolapsu. Ľudská bytosť je určitým spôsobom konfrontovaná s neľudským perfekcionizmom algoritmov AI, ktorý je na hony vzdialený od empatie, emócií a nedokonalostí človeka, na ktoré však dokáže katastrofálne pôsobiť.

Naviac, perimeter závislosti od pozitívneho hodnotenia nie je ohraničený reálnymi osobami z môjho okolia, ale rozširuje sa na tisíce až desiatky tisíc virtuálnych osôb, ktorých hodnotenie má na našu psychiku enormný dosah. Ľudský mozog sa nevyvinul tak, aby bol schopný prijímať spoločenské ohodnotenia v minútových intervaloch a od nesmierneho množstva iných subjektov. **Pod intenzívnym vplyvom sociálnych sietí a virtuálneho sveta sa ľudia stávajú závislí na svojej viditeľnej dokonalosti a neustálom prísune krátkodobých signálov odmeňovania až do tej miery, že si to spojujú s hodnotami a s pravdou. Skutočne hodnotné a pravdivé sa tak stáva to, čo prináša najviac pozitívnych hodnotení a čo ľudí udržiava v kontrolovanom pozitívnom stave krátkodobých signálov odmeňovania.**

Vzhľadom na bezprecedentnú rozšírenosť týchto sociálnych sietí (a tým aj prakticky najrozšírenejšie využívanie systémov AI, ktoré interagujú s človekom a učia sa na reálnych dátach ľudí) ide o zásahy, ktoré ovplyvňujú celú modernú spoločnosť. Častokrát sa objavuje argument, že ide o krátkodobý výkyv, ktorý sprevádzal nástup akejkoľvek pokrokovej technológie v dejinách, resp. príchod ďalšej priemyselnej technológie a veľkej spoločenskej zmeny, a že je len otázkou času, kým si na to ľudstvo privykne. Dovolíme si tvrdiť – spolu s mnohými odborníkmi na etiku sociálnych sietí a informačnú spoločnosť – že v tomto prípade je to iné a to z dvoch dôvodov.

Realitou súčasnej digitálnej éry a ostatnej priemyselnej revolúcie je jej časová neohraničenosť – nachádzame sa v období permanentnej technologickej revolúcie, ktorá nie je uzavretým procesom, ktorého dôsledky spoločnosť vstrebáva. Ľudstvo je skôr vystavené neustálemu dopingu technológií, informácií a vplyvov systémov AI, ktorého ovocím je disproporcia medzi technikou a schopnosťou spoločnosti ju správne a bezpečne využívať.

Druhým dôvodom je fakt, že informačné technológie stojace za týmito systémami sa nielen permanentne, ale predovšetkým exponenciálne zlepšujú. Na jednej strane interakcie je tak ľudský mozog, ktorý sa ďalej nevyvíja a na strane druhej technológie, ktoré sa za posledných šesťdesiat rokov mnohonásobne zmenili a porástli (len napríklad výkon počítačov sa za ten čas zvýšil v násobkoch miliárd).

V konfrontácii uvedeného sa tak dotýkame niečoho, čo sme už skôr spomenuli ako technologickú singularitu a čo je pre niektorých potvrdením obáv a výziev, ktoré obnáša koncept transhumanizmu, konkrétne vytvorením nových druhov ľudí genetickými manipuláciami alebo integráciu systémov AI a človeka pre eliminácie dôsledkov činnosti moderných systémov AI v službe sociálnych sietí a schopnosti spoločnosti ich asimilovať a integrovať.

Vyššie sme spomenuli, že nutnou súčasťou extrémne rýchleho súčasného vývoja v oblasti umelej inteligencie je fungujúca možnosť monetizácie nasadenia systémov AI. **Pri riešení dôsledkov činnosti systémov AI je treba stále pamätať na to, aké ciele sú zadané pre činnosť týchto systémov**. Implementácia algoritmov s optimalizáciou na maximalizáciu zisku sa na sociálnych sieťach premieta do predlžovania času stráveného na sieti, počtu zhliadnutých reklám, množstva dát, ktoré používateľ vyprodukuje a ktoré je následne možné nejakým spôsobom premeniť na finančný benefit.

Samozrejme, takáto optimalizácia prináša aj svoje dôsledky: sociálne bubliny, v rámci ktorých sú podsúvané len tie informácie, ktoré korešpondujú s pohľadom používateľa a ponúkanie kontaktov na osoby rovnakého razenia, uprednostňovanie falošných informácií, pretože tie viac udržia používateľa pripojeného na sieti, ergo viac zarábajú.

Pre zadané ciele dnešné sociálne siete vytvárajú prostredie (t.j. algoritmy AI takto reagujú na zadané požiadavky a vylepšujú ponúkané informácie), ktoré je toxické - odtrhnuté

od reality, s eróziou hodnôt, postmodernou[4] rezignáciou na hľadanie pravdy, zámenou dialógu za konfrontáciu,... Dôsledkom sociálnych sietí poháňaných sofistikovanými systémami AI je deštruktívne ovocie pre život a rozvoj človeka, kvalitu jeho života, vzťahy i celú spoločnosť. Jednoducho povedané – **zlyháva etický rozmer nasadenia AI**.

Kedy AI prekoná človeka (slide 28)

Mnohí odborníci, pohybujúci sa v oblasti umelej inteligencie, podvedome čakajú na moment, keď umelá inteligencia premôže ľudskú silu a inteligenciu. Inak povedané, kedy nastane vek uvedomelej AGI, ktorá nás dokáže nahradiť v práci a bude múdrejšia než my. Stále sa domnievame, že táto otázka nie je na programe dňa a aktuálne je viac súčasťou pracovnej náplne futurológov.

Skôr sa stotožňujeme s pohľadom Tristana Harrisa, bývalého etika dizajnu vo firme Google a spoluzakladateľa Centra pre humánne technológie, pre ktorého je **oveľa dôležitejší ten moment, v ktorom technológia prekoná a ovládne ľudské slabosti. Už vtedy prichádza víťazstvo AI a porážka ľudstva, lebo už vtedy prichádza závislosť, polarizácia a radikalizácia spoločnosti, zaslepenosť, strata schopnosti komunikovať a hľadať pravdu,...** jednoducho prehráva všetko ľudské v nás...

Míľnikom, ktorého by sme sa mali obávať, teda nie je budúca technologická singularita v oblasti umelej inteligencie, v ktorej AI prevýši náš intelekt, ale oveľa skôr moment, keď technológia ovládne a prekoná naše slabosti... už vtedy prichádza víťazstvo umelej inteligencie a porážka ľudstva.

Vybrané riziká generatívnych systémov (slide 29)

Generatívne systémy, či už ide o modely jazykové, grafické alebo generujúce video, vedia byť úžasné v spracúvaní obsahu zadania, generovaní odpovedí a výstupov, pričom častokrát dokážu rýchlo ponúknuť lepšie výsledky než ľudia. **Využitie je skutočne širokospektrálne a môže byť úspešné pri chápaní rizík a dôslednom aplikovaní zásad správneho použitia a kontroly.**[5]

Avšak **jedným z hlavných problémov sú nesprávne odpovede, resp. tzv. halucinovanie, t. j. vymýšľanie si odpovedí, ktoré systém predkladá ako relevantné a správne.** Treba si uvedomiť, že pri ChatGPT a ostatných generatívnych systémoch AI ide

---

4 Treba si uvedomiť, že **informačný svet, zasadený do postmoderny s jej relativizáciou pravdy, rezignáciou na racio a akcentovaním emócií a zážitkov, je jedným z kľúčov k chápaniu výziev, rizík a potenciálu, ktorý v sebe vyššie uvedené problémy obnášajú**.

5 **ChatGPT (GPT 3.5) neznamená ani tak technologický prielom ako skôr sociologický a psychologický zlom v prístupe verejnosti k systémom a možnostiam umelej inteligencie.**

v zásade o štatistické systémy, ktoré určitým spôsobom generujú čo najpravdepodobnejšie odpovede na základe naučených dát.

Generatívne systémy zo svojej podstaty nevedia, či je odpoveď správna, a či nie, keďže ponúkajú len štatisticky najpravdepodobnejšie výstupy. A tak miesto odpovede „neviem“ ponúkajú vymyslené odpovede, u ktorých neznalý používateľ nevie rozlíšiť, do akej miery sú pravdivé a do akej ide o výmysel. Preto pre úspešné nasadenie v danej oblasti je treba najprv mať jasne definované dáta, z ktorých sa učí, čo je možné v špecializovaných využitiach, a následne u všetkých generatívnych systémov odpovede aj overovať.

S problémom halucinácie je ruka v ruke spojené i riziko predsudkov a neobjektívnych výstupov. Treba však uviesť, že predsudky nie sú doménou len generatívnych systémov, ale dotýkajú sa celého spektra aj starších alebo odlišných technológií umelej inteligencie.

Ďalším problémom je napríklad nejasný spôsob narábania s údajmi, čoho dôsledkom môže byť únik dôverných dát, prehrešky voči ochrane osobných údajov i problémy s autorskými právami.

Etické riziká generatívnych systémov (slide 30)

Veľké generatívne systémy umelej inteligencie sa učia na extrémne veľkej množine dát, ktorú ponúka internet, prípadne neustály zber dát v rámci informačných systémov súčasnosti. **Systémy umelej inteligencie skutočnú inteligenciu nemajú, len ju simulujú. Takže okrem iného vôbec nedisponujú schopnosťou rozlišovať morálne dobré a zlé.**

Generované výsledky tak môžu byť veľmi problematické. Systém napríklad môže na vyžiadanie dať veľmi presný návod na samovraždu, môže generovať nelegálny obsah, tzv. deep fake systémy môžu vytvárať falošné obrazové, video a audio záznamy ľudí, napr. vytvorenie falošných pornografických materiálov takmer kohokoľvek, kompromitujúce audio nahrávky verejne činných osôb, falošné vydieranie rodičov vygenerovanými telefonátmi ich detí po fiktívnej autonehode alebo únose a podobne.

Preto sa **do súčasných generatívnych systémov implementujú viaceré úrovne kontroly**, či už ide o modifikáciu spôsobu učenia sa a generovania výstupov alebo o výstupné filtre, ktoré detegujú a blokujú problematický obsah. Každý systém je však zraniteľný, takže ide o neustály proces hľadania a implementácie čo najvhodnejších riešení.

Najvhodnejších je asi to správne slovo, lebo **to, čo slúži na ochranu, by sa dalo zneužiť i na cenzúru alebo sledovanie**.

Jedným z veľkých rizík generatívnych systémov je riziko tzv. digitálnej demencie, pri ktorej prichádza k degradácii intelektuálnych schopností a k dopamínovej závislosti na základe ponorenia sa do virtuálneho sveta, v ktorom systémy umelej inteligencie v čoraz väčšej miere suplujú kognitívne činnosti človeka, a to spôsobom, na ktorý nie sme evolučne vôbec pripravení.

Riziko digitálnej demencie (slide 31)

Osobitne môže byť tento problém vypuklý u detí a dospievajúcich. Neuvážené suplovanie viacerých kognitívnych úkonov systémami umelej inteligencie môže viesť k postupnému znižovaniu kognitívnych schopností. **Problém digitálnej demencie tak môže prispieť k narúšaniu psychického vývoja priskorým, resp. nezrelým používaním digitálnych technológií a technológií umelej inteligencie**.

Riziko digitálnej demencie sa snúbi s čoraz viac sa otvárajúcimi inteligenčnými nožnicami, keďže časť študentov, využívajúc potenciál digitálnych prostriedkov, sa bude neskutočne zlepšovať, pričom však ostatní, ktorých bude väčšina, sa budú v intenciách digitálnej demencie prepadať.

Pokiaľ nevyriešime tento deevolučný problém, budeme stáť pred veľkou civilizačnou výzvou. Použijúc Gaussovho rozdelenie môžeme povedať, že spoločnosť je natoľko vyspelá, nakoľko je vyspelá jej väčšina. **Nezávisle od vyspelej špičky by celá spoločnosť ako taká mohla upadať a nastávalo by nielen postupné znižovanie IQ, ale i úpadok celej intelektuálnej úrovne spoločnosti**.

Ľudská slabosť... (slide 32)

Viaceré vynikajúce osobnosti z oblasti umelej inteligencie hovoria o dvoch možnostiach budúceho vývoja. Ak fenomén umelej inteligencie zvládneme nielen v rovine technologickej, ale predovšetkým v rovine etickej i v celom spektre dôsledkov pre človeka a spoločnosť, pôjde o jednu z najúžasnejších vecí, s ktorými sa ľudstvo v rámci svojho civilizačného vývoja stretlo.

Ak to však nezvládneme a ontologické dôsledky nám prerastú cez hlavu, ako spoločnosť máme problém a z dlhodobého hľadiska nezadržateľne sa blížiacu civilizačnú krízu, pričom tým vôbec nie sú myslené katastrofické scenáre z oblasti sci-fi.

Povedané rečou moderného technokrata - ide o emergentnú a disruptívnu technológiu, teda novú a rozvíjajúcu sa technológiu, ktorá nahrádza a ruší technológie predošlé a úplne mení spôsob, akým spoločnosť funguje, pričom má potenciál posunúť civilizáciu ďalej.

A dodávame to, čo už technokrat nepovie - ide o technológiu, ktorú ak nezvládneme, má potenciál nás priviesť k skutočne veľkým problémom.

Príloha

Základné pojmy v oblasti umelej inteligencie (slide 34-43)