Zodpovedný prístup k umelej inteligencii

Schopnosť AI dopĺňať, ovplyvňovať či nahrádzať ľudské chápanie a rozhodovanie prináša so sebou aj množstvo nových spoločenských otázok. Nasadzovanie nástrojov umelej inteligencie nám totiž nielen otvára nové možnosti, ale aj nové spoločenské a etické riziká. Medzi takéto výzvy, s ktorými bude potrebné sa vysporiadať, môžeme zaradiť napríklad riziká neúmyselnej diskriminácie v dôsledku neférových rozhodnutí AI, alebo správna miera informovanosti dotknutých osôb o tom, že dané rozhodnutie bolo vykonané umelou inteligenciou .

Ďalším z veľkých problémom AI algoritmov je aj miera ich vysvetliteľnosti a to, že v niektorých prípadoch nevieme nahliadnuť do ich útrob tak, aby sme pochopili, ako sa rozhodujú.

Mali by sme sa pýtať nielen na to, či je daný algoritmus dostatočne efektívny v dosahovaní stanovených cieľov, ale aj na jeho potenciálne dopady, ktoré budú vyplývať z jeho rozhodnutí.

Keď používame AI na rozhodnutia, ktoré majú dopad na ľudské životy a ich kvalitu, ako zabezpečíme aby bol tento proces férový, vysvetliteľný a transparentný? Ako dokážeme identifikovať, či je rozhodnutie AI nezaťažené rôznym napr. rasovými, alebo rodovými predsudkami? Kedy môžeme zveriť proces rozhodovania do rúk AI bez toho, aby sme ju kontrolovali, a kedy bude dôležité ponechať aj dozor človeka? Ako využívať výhody plynúce z analýzy a predikcie ľudského správania pomocou AI a zároveň si nenarušiť hranice potrebné pre ľudské súkromie?

Prečo by sme mali poznať riziká AI

Ak chceme predísť rôznym nežiadúcim dôsledkom, ktoré môžu ohroziť nielen reputáciu firiem využívajúcich AI, ale aj naštrbiť dôveru spoločnosti voči AI ako takej, bude nevyhnutné, aby sme si na tieto otázky vedeli odpovedať s dostatočným predstihom. Malo by nám ísť o to využívať potenciál plynúci z nasadenia AI zodpovedne a bezpečne. Inak nám hrozí, napríklad aj to, že ľudia nebudú AI systémom plne dôverovať, nebudú ju ochotní toľko používať či rozvíjať.

Okrem toho nám nasadzovanie AI môže pomôcť zvýšiť kvalitu života nielen po materiálnej, ale aj po etickej stránke. Ak budeme vedieť posudzovať jednotlivé implementácie AI a ich etické dôsledky, pomôže nám to aj identifikovať a preferovať tie z nich, ktoré vedú k lepšiemu stavu sveta, v ktorom pracujeme a žijeme.

Tieto otázky sa musíme naviac naučiť zodpovedať my sami, tu na Slovensku. Množstvo algoritmov AI k nám prichádza zo zahraničia a môžu byť vytrénované na vzorkách, ktoré u nás nemusia fungovať hodnoverne. Ak bude napríklad výrobná linka v továrni sledovať tváre pracovníkov pri stroji a upozorňovať na ich únavu, algoritmus trénovaný povedzme v Kórei

môže byť u nás nepresný kvôli odlišným črtám tvárí na ktorých bol trénovaný, čo by mohlo viesť k nehodám.

Iné dovezené AI technológie u nás môžu po technickej stránke fungovať bezchybne, ale nebudú kompatibilné s našimi hodnotami. V Číne napríklad vyvinuli inteligentné semafory, ktoré rozpoznajú chodcov, čo prebehnú na červenú, a ich meno a tvár hanlivo zobrazia na digitálnom bilborde. Takéto niečo by u nás nepochybne narazilo na naše vnímanie (či legislatívu) ochrany súkromia, ale pri iných príkladoch to nemusí byť také jednoznačné a môže si to vyžadovať hlbšie bádanie.

Nad rozsahom možných rizík, ktoré sa týkajú AI by sme sa mali zamýšľať počas celého jej vývojového cyklu - od návrhu, cez vývoj, nasadenie až po používanie UI. Taktiež by sme nemali hovoriť len o problémoch, ktoré sú už súčasťou nášho sveta, ale analyzovať aj problémy, ktoré nás ešte len čakajú, ak budeme vo vývoji AI postupovať tak ako doteraz.

Ak je naším záujmom zvýšiť dôveru v používanie AI a znížiť mieru spoločenských a morálnych konfliktov, potom nás to vedie k potrebe zodpovedného a riadeného vývoja AI, ktorý bude schopný predikovať možné negatívne dopady a aktívne pracovať na ich znížení.

Toto všetko by sme mali vyhodnocovať a spracovať, aby sme AI mohli používať zodpovedne. Získané poznatky potom môžeme ponúkať ďalej a aj týmto spôsobom byť vo svete AI relevantným hráčom. Tak ako sa vyvíjajú AI technológie, budú napredovať aj s nimi spojené výzvy v oblasti riadenia rizík či etiky.

Je možné, že poznanie toho, čo všetko to s nami robí, bude rovnako cenné ako vývoj samotných AI technológií. Práve etická, alebo dôveryhodná AI môže byť povestnou striebornou guľkou Európy, teda špecifickou konkurenčnou výhodou v pretekoch AI, kde nám už zdá sa ušiel prvý vlak, v ktorom sedia USA a Čína.

Je v našom vlastnom záujme, aby sa technológie ako AI používali s dôverou občianskej spoločnosti. Preto potrebujeme neustále sledovať a vyhodnocovať spoločenské dopady AI a zároveň sa snažiť, aby tieto technológie boli nasadzované etickým spôsobom tak, aby z nich mal prospech čo najširší okruh ľudí. Len tak totiž žiadaný pokrok bude môcť získať podporu širokej verejnosti ako aj verejných inštitúcií.

Iniciatívy na budovanie etickej AI

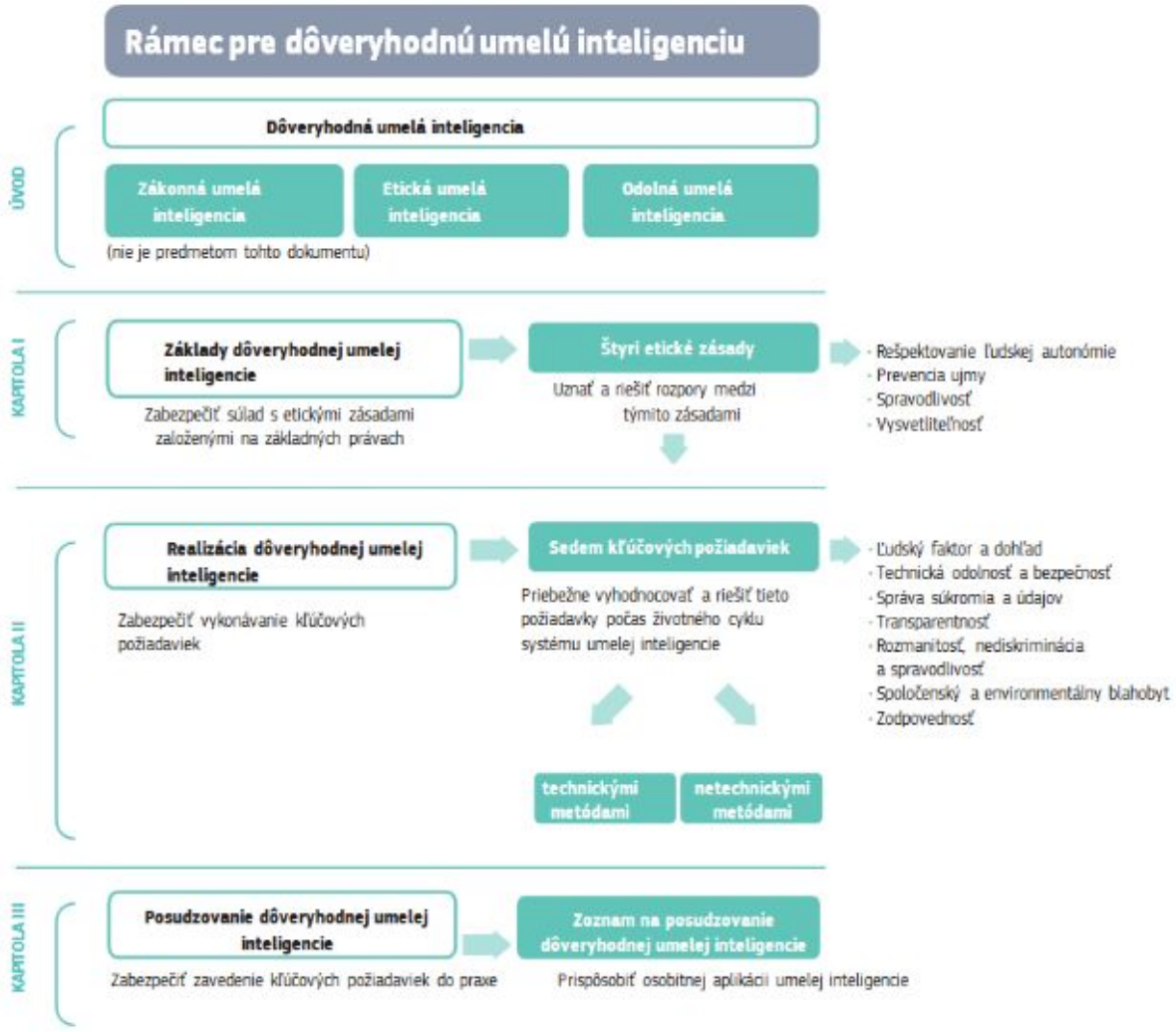
Za posledné roky vzniklo niekoľko iniciatív, ktoré sa snažia vytvoriť spoločný teoretický základ, na ktorom môžeme postupne budovať etickú AI (Asilomar AI principles - Future of Life, Montreal declaration for responsible AI, Singapur - Model AI Governance Framework , Responsible Research and Innovation, Ethical Aligned Design a mnohé ďalšie).

Tieto prístupy sa snažia o vytvorenie pravidiel a doporučení na etické princípy a hodnoty pre vývoj a nasadzovanie úzkej AI.

Medzi najvýznamnejšie patrí iniciatíva Etické usmernenia pre dôveryhodnú umelú inteligenciu z dielne expertnej skupiny na vysokej úrovni pre AI zriadená Európskou komisiou (AI HLEG) a ktorá má aj svojho slovenského zástupcu. Tieto usmernenia deklarujú pohľad na AI, ktorá bude orientovaná na človeka s cieľom rešpektovať ľudské práva a sledovania spoločenského záujmu na princípe dobrého života.

Dôležitou súčasťou tejto iniciatívy je aj pilotný program, ktorý ponúka sadu praktických požiadaviek, voči ktorým môžu vývojári AI validovať svoje vlastné riešenia a do ktorého sa môžu zapojiť aj slovenské technologické spoločnosti.

Odporúčania vychádzajú zo štyroch etických zásad (rešpektovanie ľudskej autonómie, prevencia ujmy, spravodlivosť a vysvetliteľnosť) a siedmich kľúčových požiadaviek, resp. oblastí, ktorým by sme pri vývoji AI mali venovať zvýšenú pozornosť (ľudský faktor a dohľad, technická odolnosť a bezpečnosť, správa súkromia a údajov, transparentnosť, rozmanitosť, nediskriminácia a férovosť, spoločenský a environmentálny blahobyt, zodpovednosť).



Na obrázku je návrh rámca pre etickú AI z dielne AI HLEG, ktorý vychádza z etických princípov, definuje kľúčové požiadavky a poskytuje prvý zoznam otázok na posúdenie dôveryhodnosti AI (zdroj)

Cieľom tejto kapitoly nie je poskytnúť hĺbkovú analýzu všetkých siedmich kľúčových oblastí. V nasledujúcich kapitolách chceme skôr poukázať a vysvetliť, ako niektorým oblastiam môžeme rozumieť. Mali by napomôcť k tomu, ako začať pri uvažovaní nad etickými dopadmi AI, resp. na čo nezabudnúť pri vývoji etickej AI.

Transparentnosť

Jedným zo zásadných etických problémov s AI je slabá transparentnosť a vysvetliteľnosť jej rozhodnutí. Ak umelá inteligencia pochybí, často o tom buď vôbec nevieme, alebo nedokážeme pochopiť dôvod zlyhania.

Ak má človek pocit, že mu rozhodnutie AI uškodilo, je pochopiteľné, že chce poznať dôvody, prečo, povedzme, algoritmus neodporučil, aby mu bol poskytnutý požadovaný finančný úver.

V praxi však môže program vyvinutý strojovým učením, ktorý rozhoduje o pridelení úveru, vyzerať len ako nezrozumiteľná sústava číselných váh. Banky by však mali vedieť otvorene vysvetliť, ako rozhodnutie o poskytnutí úveru vzniká.

Jeden spôsob ako to docieliť je vybrať taký AI algoritmus, ktorý sa vyznačuje práve transparentnosťou, ako sú povedzme tzv. scorecards a nepoužiť napríklad hlboké neurónové učenie, ktoré môže byť presnejšie, ale jeho správanie sa ľuďom chápe obtiažne.

Niekedy ani nie je zjavné, že sú ľudia AI algoritmom vôbec vystavení. Napríklad aj na Slovensku pribúdajú v obchodných centrách kamery, ktoré nielen chránia obchody pred krádežou tovaru, na čo sú zákazníci zvyknutí, ale aj zaznamenávajú vek, pohlavie či náladu kupujúcich, predikujú ich pohyb, alebo ich opakovane rozpoznávajú napríklad podľa bluetooth stopy mobilného telefónu.

O nasadení AI by mali organizácie otvorene komunikovať. Je tiež dôležité zvážiť, či by nemali ľudia dostať možnosť rozhodnúť sa, či vôbec chcú byť vystavení algoritmom AI a či radšej nechcú, aby sa im venoval človek. Napríklad je pozitívne, že na Slovensku pribúdajú banky, kde sa dá založiť účet prostredníctvom biometrie tváre, ale je možno rovnako dôležité, aby bola naďalej ponechaná možnosť otvoriť si účet aj klasicky na pobočke, zoči-voči bankovému úradníkovi.

Nediskriminácia a férovosť

AI algoritmy by mali každého posudzovať férovo a tak vyváženým spôsobom, aby podobné skupiny ľudí neovplyvňovali odlišne.

Preto je pri tvorbe AI algoritmov dôležité vybrať také dáta, ktoré dostatočne reprezentujú svet v ktorom žijeme, alebo aspoň tú jeho časť, v ktorej bude AI algoritmus pôsobiť.

Vybrať “reprezentatívne” dáta však nestačí. Keď napríklad v USA stroj posudzoval riziko pacientov na niektoré choroby a odporúčal preventívne prehliadky, ukázalo sa, že jeho rozhodnutia sú diskriminačné a rasovo orientované.

Algoritmus sa totiž učil a následne rozhodoval aj podľa objemu finančných zdrojov spotrebovaných v minulosti na liečbu konkrétnych pacientov. Americkí černosi však boli najmä v minulosti pri poskytovaní zdravotnej starostlivosti diskriminovaní, alebo sa sami zdráhali zdravotnícke zariadenia navštevovať a algoritmus túto zaujatosť zdedil.

Takéto predpojatosti je pred nasadením AI nutné poznať a odstrániť. Výskum v tejto oblasti je skôr v počiatkoch, už však sú k dispozícii prvé praktické nástroje schopné nájsť v algoritmoch predsudky.

Zaujímavý je aj akýsi spätný dopad nasadenia algoritmov s predsudkami - aj nám na Slovensku pomôže odhaliť podobne zakorenené diskriminácie či predsudky. Ako by u nás dopadol spomínaný algoritmus v prípade našich Rómov, ak by vychádzal z toho, koľko prostriedkov a ochoty vynakladáme na ich zdravotnú opateru oproti zvyšku populácie?

Spoľahlivosť, technická odolnosť a bezpečnosť

Zložitosť AI technológií spôsobila obavy z toho, že sa môžu zachovať nesprávne, ak sa ocitnú v nepredvídateľných okolnostiach, alebo že ich niekto s nekalým úmyslom dokáže zmanipulovať. Dôvera ľudí v umelú inteligenciu preto nezávisí len od toho, ako sa správa v bežných podmienkach, ale aj od toho, ako zvládne nečakané situácie alebo útoky.

V prvom rade je potrebné testovať algoritmy AI v najrozmanitejších podmienkach. Napríklad funkcionalitu rozpoznávania obrazu je nutné skúšať aj v prostredí so zníženou viditeľnosťou, v prítmí či v hmle.

Testy je nutné systematicky vykonávať nielen počas vývoja, ale aj nasadenia systémov. Riešenia by tiež mali obsahovať ochranu proti nežiaducim zásahom zvonka, ako sú napríklad kyberútoky.

Jeden známy typ útoku proti AI systémom na rozpoznávanie obrazu sa nazýva “adversial attack” a spočíva v dômyselnej a ľudskému oku nepovšimnuteľnej úprave obrázkov či reálnych objektov tak, že ich následne AI vidí ako niečo úplne iné.

Skupina vedcov v USA napríklad dokázala upraviť túto dopravnú značku STOP tak, že sa zmena ľuďom javí len ako akési neškodné graffiti, strojový algoritmus rozpoznávania obrazu pritom túto značku so 100 % úspešnosťou neinterpretoval ako STOPku, ale ako obmedzenie rýchlosti na 45 mph.

(zdroj)

Správa súkromia a údajov

Dôležitou oblasťou, ktorú treba spomenúť, je miera zasahovania AI do ľudského života a súkromia.

Ako príklad uveďme firmy, ktoré sa zaoberajú zberom, analýzou a vyhodnocovaním zákazníckych dát v online alebo fyzickom priestore. Ich nástroje automatizovane zbierajú a vyhodnocujú častokrát veľmi detailné údaje o správaní návštevníkov kamenného obchodu alebo e-shopu. Následne sa aj za pomoci AI snažia predikovať zákaznícke reakcie, od pravdepodobnosti návštevy obchodu, vytvorenia radov pred pokladňou až po nákup konkrétneho výrobku.

To vyvoláva hneď niekoľko otázok. Do akej miery majú firmy oprávnenie zasahovať do ľudského súkromia nielen sledovaním ale aj interpretáciou a následnou predikciou vzorcov ľudského správania za pomoci AI? Do akej miery by mali byť prevádzkovatelia týchto služieb, prípadne samotní obchodníci, úprimní voči svojim zákazníkom tým, že zverejnia informáciu o automatizovanom spracovaní dát, aj nad rámec platných právnych úprav (napr. GDPR)? Za akých podmienok máme ako dotknuté osoby nárok požiadať o to nebyť predmetom rozhodovania AI? Prípadne je možné zmysluplne kompenzovať stratu súkromia sledovaných osôb?

Extrémnym prípadom, kedy sa privátnosť dostáva do úzadia na úkor iných, prevažne ekonomických a politických záujmov, je tzv. systém sociálneho kreditu. Podobný systém by v Európe pravdepodobne narazil na silný odpor občianskej spoločnosti. Kto ale bude dohliadať na etickosť takejto technológie a jej jednotlivých funkcionalít, ak sa niektorá z firiem rozhodne podobný systém vyvinúť aj u nás?

Bude to inžinier, ktorý technológiu vyvíja, šéf spoločnosti, štátna inštitúcia alebo niekto mimo prostredia firmy, napríklad nezávislá etická komisia? A aké by mali byť profesionálne a ľudské kvality týchto arbitrov, aby sme mohli dôverovať ich rozhodnutiam?

Ľudský faktor a dohľad

Ďalšia skupina etických problémov sa viaže k rizikám autonómie AI. Musíme si stanoviť, v ktorých oblastiach a prípadoch AI dovolíme rozhodovať samostatne a kedy bude musieť byť pod dohľadom ľudí.

Príkladom riešenia je odporučenie vlády v Singapure pridať overenie človekom všade tam, kde existuje istá pravdepodobnosť pochybenia AI a zároveň je tam vysoké riziko skutočného ublíženia.

Ak povedzme AI radí, ktorú knižku si kúpiť v online kníhkupectve, je možné jej do veľkej miery prenechať voľnú ruku. Ale ak napríklad rozhoduje o chirurgickom zákroku onkologického pacienta, ľudský dozor je nevyhnutný.

Pravdepodobnosť a miera negatívneho dopadu nám môžu pomôcť zodpovedať otázku, aká miera ľudského dohľadu je pre danú implementáciu AI potrebná (zdroj)

Je nevyhnutné vyškoliť našich ľudí tak, aby chápali spôsob, akým AI funguje, poznali limity jej uplatnenia a dokázali jej rozhodnutia doplniť či zmeniť rozumným ľudským úsudkom.

Zodpovednosť

S mierou autonómie je spätá aj otázka zodpovednosti. Kto zodpovie za nehodu autonómneho auta? Bude to výrobca, používateľ AI alebo samotný stroj? Predstava právnej zodpovednosti samotného AI algoritmu vyznieva prinajmenšom zvláštne, ale rovnako sme si kedysi nevedeli predstaviť právnu zodpovednosť firiem.

Zdá sa, že napríklad v prípade plne autonómnych automobilov budeme musieť rozlišovať medzi osobou zodpovednou (reliable) za nehodu, osobou, ktorá bude braná na zodpovednosť (accountable) a osobou, ktorá bude naprávať vzniknuté negatívne dopady (liable).

Podobne ako pri iných produktoch a technológiách, aj v prípade umelej inteligencie je nutné, aby ľudia, ktorí ju vyvíjajú a nasadzujú, zodpovedali za to, čo tieto systémy robia. Ak sa máme baviť o tom, čo znamená byť za niečo zodpovedný, treba si najskôr definovať samotnú zodpovednosť. Morálna zodpovednosť za čin pochádza okrem iného aj z našej schopnosti konať slobodne. To znamená, že musí ísť o vnútorné rozhodnutie, nie rozhodnutie nanútené zvonka. Okrem toho by si mal byť subjekt vedomý toho, čo koná a byť si vedomý dôsledkov svojej činnosti.

Súčasnej AI ešte takto chápanú vedomú činnosť prisúdiť nedokážeme. Nateraz sa preto chápe len ako nástroj, i keď v niektorých ohľadoch nesmierne silný. Tento nástroj sme vytvorili my ľudia a ako tvorcovia zaň aj musíme vziať adekvátny diel zodpovednosti. Nesmieme však zmiešavať dokopy rôzne chápania zodpovednosti, ktoré sme opísali vyššie.

Systematické a odborné posudzovanie AI rizík a ich správa

Okrem už spomínaných etických usmernení z dielne AI HLEG pre dôveryhodnú AI môžeme odporučiť niekoľko ďalších praktických postupov, ktoré slúžia na systematické posúdenie rizík spojených s nasadzovaním algoritmov AI. Jedna z metodík pre vyhodnocovanie etického rizika AI bola vytvorená výskumníkmi z Johns Hopkins University a mestským zastupiteľstvom v San Franciscu, (https://ethicstoolkit.ai/) a do nášho jazyka ju preložila a verejnosti voľne sprístupnila slovenská firma exe.

Spomínaná príručka kladie sériu otázok, ktorými je možné posúdiť spôsob, závažnosť, či škálu dopadu algoritmov na našu spoločnosť. Následne skúma vhodnosť či rizikovosť použitých dát a dokáže identifikovať prípady, kedy je potrebné nastaviť kroky na mitigáciu daného etického rizika. Metodika príručky tiež napríklad posudzuje, či je dostatočne jasné, kto je za použitie algoritmu zodpovedný.

Spoliehať sa, že len samotné IT firmy budú schopné odborne identifikovať, pokryť a vyriešiť etické problémy AI, je príliš optimistické očakávanie. Naliehavo potrebujeme novú generáciu

odborníkov na etiku AI, ktorí budú rozumieť nielen filozofickej povahe problému, ale aj digitálnym technológiám. A to do takého detailu, aby dokázali správne posúdiť eticky sporné situácie a vybrať vhodné nástroje na ich riešenie.

V rámci podpory etickej AI naprieč celou firmou vzniká nielen potreba doplniť kompetencie aktuálnych rolí vo firme, ale aj vytvorenie pozície Ethics officer, či Trust officer. Títo ľudia by mali mať na starosti práve výklad morálnych intuícií a výber vhodných nástrojov, ktoré sa použijú pri vývoji etickej a dôveryhodnej AI.

Príklad, ako sa môže prejaviť implementácia etickej AI do kompetencií aktuálnych, prípadne nových rolí vo firme

Na záver je potrebné dodať, že sme si s plnou pokorou vedomí, že máme pred sebou veľa práce a veľa toho, čo sa ešte musíme naučiť. Slovensko je aktuálne pozadu nielen čo sa týka rozvoja AI, ale aj jej etickej a spoločenskej reflexie. Myslíme si, že čas ignorovať etické problémy nových digitálnych technológií, ako je umelá inteligencia, sa blíži aj na Slovensku k svojmu koncu. V skutočnosti vypršal práve teraz.