Asistívna umelá inteligencia a ľudské hodnoty autonómie, úsilia a rozmanitosti

Martin Takáč

Centrum pre kognitívnu vedu FMFI UK
Mlynská dolina, 848 48 Bratislava
Email: martin.takac@fmph.uniba.sk

Abstrakt

V súčasnosti sa rozmáha trend rôznych asistujúcich aplikácii na báze umelej inteligencie (UI). Aplikácie zbierajú o používateľovi údaje, následne na ich základe budujú predikčný model používateľa a asistujú mu v rôznych činnostiach, prípadne nad nimi priamo preberajú kontrolu. Okrem etickej otázky kto zbiera dáta, čo z nich vie vyčítať a načo ich použije, sú tu aj ďalšie: ako asistívna UI ovplyvňuje ľudskú slobodu rozhodovania? Ako tvaruje naše voľby a naše správanie? Človek je komplexný tvor, ktorý nie je len racionálny optimalizátor, ale má aj iné aspekty, ktoré nemusia byť modelovateľné a predikovateľné. Ak človeka modelujeme ako mechanizmus a tak s ním interagujeme, nemechanizovateľné aspekty marginalizujeme a tým jednotlivca aj spoločnosť istým spôsobom tvarujeme. V mojom príspevku sa okrem tohto aspektu zameriam na vzťah asistívnej UI k hodnote ľudského úsilia, strádania a emočného vkladu a tiež poukážem na evolučné riziko koncentrácie, unifikácie a globalizácie technologických riešení a ich vzťah k hodnote rozmanitosti.

1 Úvod

V súčasnosti sa rozmáha trend rôznych asistujúcich aplikácii na báze umelej inteligencie (UI). Aplikácie zbierajú o používateľovi údaje, následne na ich základe budujú predikčný model používateľa a asistujú mu v rôznych činnostiach, prípadne nad nimi priamo preberajú kontrolu. Samotné zbieranie a ukladanie dát a ich následné využitie je eticky citlivá téma a venujú sa jej napr. Zuboff (2019) alebo Stephens-Davidowitz (2017). V tomto príspevku sa skôr zameriam na to, ako asistívne aplikácie môžu ovplyvniť ľudskú autonómnosť (Časť 2), teda ako tvarujú naše voľby a naše správanie a z dlhodobého hľadiska celú spoločnosť. Následne sa zamyslím nad hodnotou ľudského úsilia a ako náš postoj k nej ovplyvňujú asistívne technológie (Časť 3). V závere dám do kontrastu spôsob globalizovaného nasadzovania technologických riešení a evolučné procesy podporujúce rozmanitosť a paralelné pomalé testovanie viacerých riešení.

2 Hodnota ľudskej autonómnosti

Danaher (2019) definuje tri podmienky nutné pre autonómne rozhodovanie: (1) mentálnu kapacitu robiť rozhodnutia, (2) dostupnosť adekvátnych možností výberu, a (3) voľbu bez nútenia a manipulácie. Asistívne technológie nám, najmä v neprehľadnej džungli informácií dostupných na internete, môžu urobiť predvýber a ukážu nám možnosti, ktoré by sme kvôli nadbytku šumu mohli prehliadnuť. Na druhej strane predvýber naše možnosti limituje. Danaher (2019) delí algoritmické nástroje podľa stupňa obmedzovania možností nasledovne:

Nástroje s človekom-definovanými preferenciami: človek presne špecifikuje, čo chce, aby algoritmus urobil, a algoritmus mu pomáha dosiahnuť tento výsledok;

Nástroje s výberom preferencií z menu: človek nešpecifikuje svoj preferovaný výsledok, ale vyberá si z ponuky možností, ktoré mu poskytne algoritmus;

Nástroje s predpovedanými preferenciami: algoritmus sa na základe dolovania údajov (často od iných ľudí) snaží predpovedať, čo bude používateľ chcieť, a podľa toho vytvára možnosti;

Nástroje so samo-obmedzujúcimi preferenciami: algoritmus funguje ako nástroj predbežného záväzku, ktorý uprednostňuje dlhodobé záujmy používateľa (možno uvedené, možno predpovedané) pred jeho okamžitými záujmami.

Okrem okamžitého vplyvu na autonómiu rozhodovania sa však treba zamyslieť aj nad dlhodobým vplyvom na spoločnosť. Čo sa deje s našimi životmi, v spoločnosti, ktorá integrovala veľké dáta, prediktívnu analytiku a inteligentné prostredia? Najmä s nástupom masového využívania generatívnej umelej inteligencie, či už v podobe četbotov na báze veľkých jazykových modelov, rôznych kopilotov, nástrojov na sumarizáciu obsahu a odporúčacích systémov, začína dochádzať k akejsi technofúzii—prelínaniu naších kognitívnych schopností so schopnosťami našich inteligentných nástrojov. Frischmann a Selinger (2018)

v knihe *Re-Engineering Humanity* upozorňujú na to, že cieľ dizajnovať predikovateľné a programovateľné svety ide ruka v ruke s tvarovaním predikovateľných a programovateľných ľudí. Nemusí ísť o vedomý zámer, ale môže to byť nezamýšľaný dôsledok. Človek je komplexný tvor, ktorý nie je len racionálny optimalizátor, ale má aj iné aspekty, ktoré nemusia byť modelovateľné a predikovateľné.[1] Ak človeka modelujeme ako mechanizmus a tak s ním interagujeme, nemechanizovateľné aspekty marginalizujeme a tým tvarujeme spoločnosť a jednotlivca istým spôsobom. Jedným z nebezpečenstiev je, že sa sami začneme vnímať ako nástroje, ktoré sa majú upravovať, optimalizovať a programovať, akoby naše mysle a telá neboli ničím iným ako technológiami. Autori v spomínanej knihe hovoria, že našou najväčšou výzvou je otázka ako využívať technológie bez toho, aby sme sa nimi sami stali.

Leonhard (2018) vyzdvihuje hodnotu ľudskej pomalosti a omylnosti: to, čo nás robí šťastnými—krásu, umenie, neočakávané maličkosti, zdanlivo iracionálne akcie, nedosiahneme, ak našim hlavným cieľom bude optimalizovať, racionalizovať a urýchľovať ľudské správanie. Umelá inteligencia nás môže odbremeniť od nudných činností, ale nemusí nutne vytvoriť čas a priestor pre to, čo nás napĺňa a robí šťastnými. Ako (údajne) povedal Martin Heidegger, „Technika prekonala všetky vzdialenosti, ale nevytvorila nijakú blízkosť."

V tejto súvislosti stojí za hodno spomenúť aj prácu psychiatra Iaina McGilchrista *The Master and His Emissary: The Divided Brain and the Making of the Western World* (McGilchrist, 2009). Autor sa v knihe zaoberá funkčnou asymetriou mozgových hemisfér. Poukazuje na evolučný význam dvoch typov pozornosti: fokusovanej, dôležitej najmä na nájdenie či ulovenie potravy, a difúznej, holistickej, dôležitej na detekciu potenciálneho ohrozenia predátormi. V kontexte hemisfér by ľavostranné spracovanie zabezpečovalo fokus, orientovanosť na cieľ, manipuláciu, redukcionizmus, istotu, a re-prezentáciu už poznaného. Pravostranné spracovanie je orientované na spojenie, vzťah, exploráciu možností, holizmus a prítomný okamih. McGilchrist tvrdí, že v súčasnej západnej kultúre historicky začal nezdravo dominovať „ľavostranný" prístup k svetu na rozdiel od celostného vnímania. Symptómom tejto dominancie môže byť aj technooptimizmus, ktorý dúfa, že na problémy ľudstva stačí nájsť technologické riešenia. Vyššie popísaný trend „mechanizácie" ľudstva by túto dominanciu ďalej podporoval.

[1] Ďakujem Tomášovi Gálovi za upozornenie, že existencia takýchto aspektov je mojím predpokladom, ktorý čaká na empirické overenie. K tejto otázke viď diskusiu B. Skinnera a C. Rogersa v Kirschenbaum a Henderson (2016) a prácu Takáč (2020).

3 Hodnota ľudského úsilia

Jedným z dlhodobých problémov ľudstva je duševné zdravie. Odhadom 1 z 3 žien a 1 z 5 mužov počas svojho života zažije ťažkú depresiu (Dattani a spol., 2023). 4,4% celosvetovej populácie trpí úzkosťami a 4% depresiami (Dattani a spol., 2023). S nástupom veľkých jazykových modelov a na nich založených četbotov sa skúmajú možnosti psychoterapie sprostredkovanej inteligentnými aplikáciami, ako Woebot (Fitzpatrick a spol., 2017), Wysa, Elomia, Mindspa (Haque a Rubya, 2023). To je dôležité najmä v kontextoch, kde je psychoterapia nedostupná, či už kvôli vyťaženosti psychoterapeutov alebo cenovej nedostupnosti psychoterapie pre isté sociálne skupiny (Macháčková a Dostál, 2022). Ukazuje sa, že jedným z faktorov účinnosti psychoterapie je klientom vnímaná empatia (Petrovická, 2022), preto sa skúma, ktoré verbálne a neverbálne signály takéto vnímanie generujú (Loveys, 2021; Loveys a spol., 2022). Zavádza sa pojem *pragmatická autenticita*, ktorá na rozdiel od skutočnej, teda *ontologickej autenticity* vytvára interakcie vnímané používateľmi ako skutočné a zmysluplné bez ohľadu na to, či agent disponuje replikovanými ľudskými procesmi autenticity, empatie, atď. (Geld, 2024).

Mimo psychoterapeutického kontextu sa ponúka možnosť využiť služby četbotov na zástupné napísanie emočne podfarbených správ pre ľudí v ťažkých situáciách, napr. kondolenčných listov, alebo jednoducho ľuďom, s ktorými máme citový vzťah. Experimenty však ukazujú, že ľudia negatívne vnímajú, ak ich priateľom s písaním citovo podfarbených správ pomáhal nástroj umelej inteligencie (Kirk a Givi, 2025). Vnímajú to ako nedostatok citovej investície zo strany ich priateľa (Liu a spol., 2024).

To, čo potrebujeme a oceňujeme v krízových situáciách, je autentický ľudský záujem. Vedomie, že ten druhý vložil do komunikácie emočné úsilie, že možno musel hľadať slová, prekonávať ťažkosti, dotýkať sa našej bolesti. Neplatí to však vždy: Ľudia, ktorí interagujú s programom Replica[2], vedia, že Replica nie je živý človek. Napriek tomu mnohí tvrdia, že im efektívne nahrádza ľudského romantického partnera (Singh-Kurtz, 2023). Takéto vzťahy sú, na rozdiel od vzťahov s reálnymi ľuďmi kvôli submisívnosti UI partnera inherentne asymetrické a netvoria model zdravých partnerských vzťahov. U aplikácii typu Replika navyše existuje riziko vytvárania závislosti, najmä u zraniteľných skupín používateľov.

4 Hodnota rozmanitosti

V poslednej časti článku sa zamyslím nad rizikami spojenými s globálnym nasadzovaním malého počtu riešení resp. inteligentných nástrojov. 19. júla 2024 ovplyv-

[2] `https://replika.com`

nila chybná aktualizácia programu vydaná kyberbezpečnostnou firmou CrowdStrike milióny zákazníkov so systémom Microsoft Windows po celom svete. Výpadok, ktorý trval 72 hodín, výrazne narušil aj letovú dopravu: z 411 009 celosvetovo plánovaných letov muselo byť zrušených 16 896, čo predstavuje niečo vyše 4% (Hetzel, 2024). Priame škody spôsobené leteckým spoločnostiam skupiny Fortune 500 boli vyčíslené na 5,4 miliardy amerických dolárov (Fung, 2024).

O'Neil (2016) v knihe *Weapons of math destruction* píše, že moderné inteligentné nástroje sú nebezpečné kvôli kombinácii troch faktorov: netransparentnosti procesu rozhodovania, schopnosti výrazne zasahovať do ľudského života (napr. UI systémy na prideľovanie hypoték a pôžičiek, prijímania do školy či do zamestnania) a ich masovému nasadeniu: pokiaľ vás systém vyradí spomedzi uchádzačov o zamestnanie v jednej firme, môžete to skúsiť inde. Pokiaľ by však väčšina firiem používala rovnaký UI systém, vyradí vás globálne.

Ďalším rizikom je koncentrácia moci, ktorú prístup k dátam a poskytovanie globalizovanej UI služby vytvára: či už sú to jednotlivci ako Elon Musk, alebo spoločnosti ako Google, možnosť zneužitia je obrovská.

Najnovší rozmach umelej inteligencie vzbudzuje obavy, že môže zasiahnuť do prirodzeného evolučného procesu vývoja druhov a destabilizovať prírodnú rovnováhu (Vavruška, 2024). Evolúcia je pomalá, postupuje metódou pokus—omyl a vďaka rozmanitosti sú vždy paralelne skúšané viaceré riešenia, čo vytvára potrebnú robustnosť. V evolúcii nejde len o efektivitu, dobrým výsledkom nie je optimum, ale rovnováha. Racionálne vykalkulované riešenia nemusia na prežitie ľudského druhu stačiť. Vavruška preto navrhuje podriadiť UI prirodzeným evolučným záujmom ľudstva. Prichádza s deklaráciou tzv. „AI for people" (AI 4P), ktorá:

- podporuje zdravé fungovanie ľudského mozgu,
- nepodkopáva vedúcu úlohu ľudskej civilizácie,
- neohrozuje spoločnosť dátovou centralizáciou,
- posilňuje slobodnú vôľu ľudí,
- nevnucuje dogmy,
- poskytuje lojálne služby jednotlivcom,
- neohrozí zmysel života a evolúcie,
- podporuje kultúrnu toleranciu (Vavruška, 2024).

Zároveň by podľa deklarácie AI 4P isté typy UI označené ako „strategická UI" mali byť pod štátnou a medzinárodnou kontrolou.

5 Zhrnutie

V článku sme vychádzali z hodnotovo orientovaného prístupu. Analyzovali sme možné vplyvy asistívnej umelej inteligencie na ľudské hodnoty autonómie, autentického úsilia a rozmanitosti. Poukázali sme na to, že nekritické aplikovanie mechanistického modelovania ľudského správania marginalizuje tie stránky jednotlivca a ľudskej spoločnosti, ktoré nie sú mechanistické. Zároveň sme načrtli prístup, ktorý sa stavia za evolučný záujem ľudstva a podriaďuje mu ďalší vývoj inteligentných technológií.

Poďakovanie

Výskum bol podporený grantom VEGA 1/0373/23 a KEGA 022UK-4/2023.

Literatúra

Danaher, J. (2019). The ethics of algorithmic outsourcing in everyday life. Yeung, K. a Lodge, M. (zost.), V *Algorithmic Regulation*. Oxford University Press.

Dattani, S., Rodés-Guirao, L., Ritchie, H. a Roser, M. (2023). Mental health. *Our World in Data*.

Fitzpatrick, K. K., Darcy, A. a Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19.

Frischmann, B. a Selinger, E. (2018). *Re-Engineering Humanity*. Cambridge University Press.

Fung, B. (2024). We finally know what caused the global tech outage - and how much it cost. https://edition.cnn.com/2024/07/24/tech/crowdstrike-outage-cost-cause/index.html July 24, 2024 (Accessed on 2025-04-18).

Geld, B. (2024). Faking it right: Human-likeness in pragmatic authenticity. Diplomová práca, University of Vienna.

Haque, M. D. R. a Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR Mhealth Uhealth*, 11:e44838.

Hetzel, J. (2024). The crowdstrike it outage: What does it mean for the airline industry? https://www.cirium.com/thoughtcloud/crowdstrike-it-outage-what-does-it-mean-for-airline-industry Oct 7, 2024 (Accessed on 2025-04-18).

Kirk, C. P. a Givi, J. (2025). The ai-authorship effect: Understanding authenticity, moral disgust, and consumer responses to ai-generated marketing communications. *Journal of Business Research*, 186:114984.

Kirschenbaum, H. a Henderson, V. L. (2016). *Rozhovory s Carlem R. Rogersem*. Portál.

Leonhard, G. (2018). *Technológia a humanita*. SIEA.

Liu, B., Kang, J. a Wei, L. (2024). Artificial intelligence and perceived effort in relationship maintenance: Effects on relationship satisfaction and uncertainty. *Journal of Social and Personal Relationships*, 41(5):1232–1252.

Loveys, K. (2021). *Developing Engaging Digital Humans for Psychotherapeutic Applications*. Dizertačná práca, University of Auckland.

Loveys, K., Hiko, C., Sagar, M., Zhang, X. a Broadbent, E. (2022). "i felt her company": A qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. *International Journal of Human-Computer Studies*, 160:102771.

Macháčková, L. a Dostál, D. (2022). Vliv vybraných charakteristik jedince prožívajícího krizi na ochotu komunikovat s chatbotem. Šejnová, G., Vavrečka, M. a Hvorecký, J. (zost.), V *Kognice a umělý život XX.*, str. 66–73. České vysoké učení technické v Praze.

McGilchrist, I. (2009). *The Master and His Emissary: The Divided Brain and the Making of the Western World*. Yale University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Books.

Petrovická, K. (2022). Is none treatment for mental health problems better than a controversial one? Šejnová, G., Vavrečka, M. a Hvorecký, J. (zost.), V *Kognice a umělý život XX.*, str. 170–172. České vysoké učení technické v Praze.

Singh-Kurtz, S. (2023). The man of your dreams for $300, Replika sells an AI companion who will never die, argue, or cheat—until his algorithm is updated. https://www.thecut.com/article/ai-artificial-intelligence-chatbot-replika-boyfriend.html March 10, 2023 (Accessed on 2025-04-18).

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. Dey Street Books.

Takáč, M. (2020). *Myseľ ako objekt*. Absynt-Kalligram.

Vavruška, D. (2024). *Život v době robotů*. Grada Publishing.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs.