Rozhovor

Rozhovor s Jurajem Hvoreckým o umělé inteligenci, obavách z jejího rozvoje, náboženství a pečovatelských robotech.

Mgr. Juraj Hvorecký, Ph.D. je slovenský filosof působící v České republice. Studoval filosofii na Univerzitě Komenského v Bratislavě a na Slippery Rock University v Pensylvánii. Doktorát z filosofie získal na Filozofickém ústavu v Bratislavě v roce 2005. V současnosti působí jako vedoucí Oddělení aplikované filosofie a etiky na Filosofickém ústavu AV ČR. Je autorem řady odborných publikací především z oblasti filosofie mysli a aplikované etiky, zejména etiky nových technologií a institucionální etiky. Aktivně se věnuje také popularizaci společenských věd a je členem Bioetické komise Rady pro výzkum, vývoj a inovace (RVVI) a Etické komise pro posuzování otázek spojených s provozem automatizovaných a autonomních vozidel při Ministerstvu dopravy (MD).

Koncem roku 2022 zpřístupnila firma OpenAI bezplatně ChatGPT (GPT-3.5). Během pouhých dvou měsíců si tato aplikace získala přes 100 milionů uživatelů a téma umělé inteligence (AI) se rázem stalo předmětem širokých diskusí nejen v odborných kruzích, ale i mezi laickou veřejností. A podobně jako v případě covidu-19 či konfliktu na Ukrajině se jako mávnutím kouzelného proutku přes noc objevilo nepřeberné množství nových „odborníků" i na toto módní téma. My se však známe již řadu let ze setkávání při různých příležitostech, především z Centra Karla Čapka pro studium hodnot ve vědě a technice a z etických komisí při RVVI a MD. Dobře vím, že na rozdíl od mnoha jiných „mluvících hlav" je Tvůj zájem o umělou inteligenci daleko hlubší a dlouhodobější, a navíc do této diskuse vnášíš celou řadu aspektů, které bývají často opomíjeny, například pohledy z kognitivní vědy, filosofie mysli a společenských věd. Mohl bys nám tedy, Juraji, krátce představit svou cestu k tématu AI? Jak ses k němu vlastně dostal?

V současnosti mě mnoho lidí zná jako vedoucího Oddělení aplikované filosofie a etiky na Filosofickém ústavu AV ČR, ale zajímavější je samozřejmě cesta, jak jsem se na toto místo dopracoval, než pozice samotná. Jsem dítě přelomu, vysokou školu jsem začal studovat na začátku 90. let a po několika semestrech na Univerzitě Komenského v Bratislavě jsem s notnou dávkou štěstí vyhrál

stipendium do USA, kde jsem na malé *college* získal bakalářský titul. Následně po návratu na Slovensko se mi ještě díky Fulbrightovu stipendiu naskytla příležitost odjet na studijní pobyt mezi filosofické hvězdy v arizonském Tusconu. Nemohl jsem si přát víc: být formován předními světovými autoritami na epistemologii a filosofii mysli, jako jsou Alvin Goldman nebo David Chalmers, přiměje k intelektuální skromnosti, ale zároveň otevírá horizonty, které jinde nenajdete.
Má dizertace a většina další práce po příchodu do Česka (za což patří velké díky Ivanu M. Havlovi, 1938–2021) byla z oblasti filosofie psychologie a za samozřejmou jsem považoval i spolupráci s kolegy z kognitivní vědy, která tehdy byla u nás ještě v plenkách. Zajímaly mě emoce, racionalita, intencionalita, ale i teoretická psychologická témata jako modularita nebo vrozenost jazyka. Součástí tohoto prostředí byli vždy i lidé z oblasti informatiky a robotiky, a tak ani technologie mi nebyly cizí. A když se spustil *hype* kolem umělé inteligence, nebyl jsem zaskočen, jelikož jsem byl s termíny jako perceptron, učení s posilováním nebo vstupní neurony dlouho před tím dobře obeznámen.
Když Filosofický ústav začal klást větší důraz na generování aplikovaných výstupů a nabídl nám založení nového dedikovaného oddělení, byla to výzva. Podařil se nám ale husarský kousek. První rok fungování oddělení jsme věnovali psaní grantu při programu Horizont Evropa a hned napoprvé jsme jej získali, díky čemuž vzniklo CETE-P, tedy Centrum pro environmentální a technologickou etiku – Praha. Ještě bych chtěl poznamenat, že s aplikací to myslíme opravdu vážně a aktivně se zapojujeme do rozhodovacích procesů několika ústředních orgánů státní správy z pozice poradců nebo členů grémií.

Aktuálním tématem Tvého výzkumu je zejména etika umělé inteligence a příbuzná témata. Jak vnímáš diskuse o existenční hrozbě ze strany AI a předpovědi, jako jsou ty od Eliezera Yudkowského a Natea Soarese, že závod o vývoj obecné umělé inteligence nedopadne dobře, jak ostatně hlásá i titul jejich knihy z roku 2025 „If Anyone Builds It, Everyone Dies“?

Nejsem sám, kdo říká, že povaha diskuze je pomýlená: i když počet příspěvků o existenční hrozbě ze strany AI klesá (pak se ale objeví zpráva o novém jazykovém modelu Mythos od firmy Anthropic a hned jsou toho plné noviny), pořád platí, že pro stromy nevidíme les. Soustředíme se na jednotlivé excesy, projektujeme je dál do budoucna a inferujeme z nich fantaskní představy o konci lidstva. Mnohem důležitější je ale pozorně sledovat, co s námi AI dělá už dnes: jak ovlivňuje veřejnou diskuzi (znovuzavedením slov, které už ze slovníku pomalu mizely, nebo šířením dezinformací), jak kazí povahu vědecké práce (do metastudií se dostávají smyšlené práce a devalvují celkový výsledek), nebo ničí lidskou duši (sebepotvrzováním si vlastních postojů pomocí podbízivých chatbotů). To jsou skutečné etické a sociální výzvy doby. Čeká nás mnohem více práce pochopit, co se děje kolem nás tady a teď, než co může superinteligence nebo vzpoura robotů přinést ve vzdálenější budoucnosti.

Když se tedy soustředíme na dnešek, jak se díváš na problematiku stagnace AI? Někteří výzkumníci totiž naopak mluví o tom, že jsme již narazili na výkonnostní strop a vlna zájmu o AI brzy opadne.

Tady bych rozlišil dvě roviny. Stagnace z pohledu uživatele nedává smysl. Nové modely se objevují každý týden, konkurence je obrovská, a když se například podívám na výsledky AI v oblasti generování videa, pokrok za poslední rok je neuvěřitelný (což má mimochodem zásadní vliv na

epistemologii a etiku v oblasti deepfakes). Je pořád co vylepšovat, daří se nám lépe pracovat se syntetickými daty, algoritmy se optimalizují, modely zmenšují. Jenže pak je tady abstraktnější rovina: architektura transformerů, na kterých velké jazykové modely stojí, má své jasné limity a ukazuje se, že problémů jako halucinace nebo absence uvažování se nezbavíme, jelikož přímo plynou z podoby této architektury. Podobně se začíná ukazovat, že podbízivost modelů (označovaná jako *sycophancy*) není jenom výsledkem komerčních zájmů výrobců udržet nás u svého produktu, ale souvisí s povahou posilování modelů během tréninku. Opět je to něco, z čeho neexistuje jednoduchá cesta ven. Navíc, a na to se často zapomíná, „nemateriální" inteligence se opírá o fyzický základ datacenter. Narážíme na nedostatek energie k jejímu provozu, vody ke chlazení a surovin k výrobě komponentů. Stagnace tedy může přijít i proto, že nápady na zlepšení bychom měli, ale není cesta, jak je realizovat.

Přejděme nyní k tématu, které je v mimořádně sekulární a nenáboženské společnosti ČR velmi málo reflektováno, je však blízké tomuto časopisu (a věnují se mu i některé přijaté příspěvky). Co umělá inteligence a náboženství? Jak vidíš toto spojení?

Jako věřící jsem byl prvotně nadšen, že budu svou víru konzultovat s chatbotem. I mnoho kněží asi tajně doufalo, že při psaní kázání jim budou asistovat jazykové modely. Jenže pokud jde o fakta (a jako luterán si zakládám na přesném znění svatých textů snad ještě více než ostatní konfese), tyto modely je prostě nemají. Jsou to jenom generátory textu na základě pravděpodobnostních vztahů v trénovacích datech. Pravděpodobnost a jistota jsou dosti odlišné koncepty a víra je zejména o jistotě. Také jsem silně proti zbožšťování takovýchto systémů, jak se to někdy děje. Dokonalost opravdu není vlastnost, kterou by obdobné systémy disponovaly, a obávám se, že právě nešťastné úvahy o superinteligenci (tj. AI, která přesahuje lidskou inteligenci natolik, že může samu sebe zlepšovat), mohou vést i k snahám o zbožštění umělých systémů. Se zájmem ale sleduji religionistickou práci, která se snaží postihnout, jakou konkrétní formu jednotlivých náboženství jazykové modely zachycují, nebo jak formují soudobé podoby víry.
Také je důležité si uvědomit, že technologie a náboženství na sebe vždy nějaký vliv budou mít. I procesy v církvi se automatizují, dobrý hypertextový vyhledávač je nezbytností nejen pro čtenáře Bible. A to nemluvím o online přenosech nebo přesných překladových nástrojích. K nějaké hlubší syntéze to ale zatím nespěje.

Jaké další etické problémy vyvstávají z používání nových technologií?

Upozornil bych na dvě témata, o kterých se mluví méně, než by bylo záhodno. První se týká už existující a široce rozšířené technologie *smart homes*. Tyto nástroje totiž přinášejí nová rizika a nové společenské dělení. Na jedné straně se ukazuje, že výraznější úsporu času ani více prostoru pro společné sdílení do domácnosti nepřinášejí. To ale není nic nového, existuje mnoho studií, které ukazují, jak zavedení nové technologie slibující osvobodit člověka od práce mu práci naopak přidává. Jenže v tomto případě se uvedený efekt propojuje se silnou uživatelskou jednostranností – obvykle má jenom jeden partner opravdovou kontrolu nad systémem (obvykle muž), on je ten znalý a kompetentní a taková epistemická asymetrie vede k nerovnostem a potenciálně k násilí. Známe situace, kdy partner i po rozchodu nadále ovládá domovní systém a brání v legitimním užívání bydlení.
Druhá oblast, která se týká spíše blízké budoucnosti, je nasazení pečovatelských robotů a dalších

sofistikovaných asistivních technologií. Důvody pro jejich nasazení jsou nasnadě. Populace stárne a o starší a nemocné se nemá kdo starat. Navíc je to práce vyčerpávající, často vede k vyhoření, je špatně placená a genderově nevyvážená. Tak ji zrobotizujeme. Jenže tenhle postup je mimořádně diskutabilní. Není ani tak otázkou, zda roboti vůbec mohou o někoho pečovat (i když i ta se hodně diskutuje). Mně zajímá zdánlivě mnohem přízemnější diskuze: jak víme, že roboti jsou již na plnění pečovatelských úkolů připraveni? Jinak řečeno, co to znamená testovat je a uvádět do praxe? Když se podíváte na aktuální situaci, vývoj probíhá téměř výhradně v akademickém prostředí, a když komerčně, tak vždy s podporou nějakých grantů. Jenže jakmile granty skončí, nemá na vývoj kdo navázat – a je celkem zjevné, že plně funkčního robota nevytvoříte za pár let. Dále je tady zatím eticky neprozkoumaná otázka testování. Co to znamená testovat pečovatelského robota? Protože je testovací, nechcete jej zkoušet na zranitelných lidech z cílové skupiny. U nich totiž každé technické selhání (a že jich bude hodně!) znamená významné etické selhání. Tak je testujete na studentské populaci, která leccos vydrží a ráda zkouší nové věci. Jenže nakolik jsou pak získaná zjištění validní pro skutečnou cílovou skupinu? Starší lidé mají mnohem menší fyzickou výdrž, jejich mluva je často nejasná, jsou pomalí. Vytrénovat robota na populaci, která tyhle komplikace nemá, se zdá být úplně zbytečné. Jeho případné nasazení v reálném provozu pak přímo neetické. Jak trefně poznamenala jedna japonská ošetřovatelka, která má s roboty své zkušenosti: „když jej tady máme, je to pro nás jako mít dalšího pacienta s demencí“. Její poznámka otevírá mnohem širší filosoficko-sociální problém: jak pracovat s marginalizovanými skupinami, které technologie činí ještě více zranitelnými?

Srdečně Ti, Juraji, děkuji za mimořádně podnětný rozhovor, který dle mého názoru výborně uvádí čtenáře do tematiky tohoto speciálního čísla „Umělá inteligence a proměna společnosti: Reflexe současných perspektiv a budoucích výzev pro společenské vědy a humanitní obory“.

Daniel D. Novotný
koeditor *Caritas et veritas*