# Petr Špecián: Zavítal jsem na přednášky z filozofie a trajektorie mojí kariéry se navždy změnila

**„Jedeme po několika kolejích zároveň a na každé jedeme jinak rychle," popisuje současnou technologickou situaci Petr Špecián, zakladatel** AI Institutional Transformation Research Group **. Ve svých výzkumech se zajímá o adaptaci institucí na nové technologie, velké jazykové modely a jejich využití v praxi nebo obecně o podobu a vztah demokracie a expertízy v digitální době. Jak se mění výzkumné metody? Halucinují velké jazykové modely? A jaké hodnoty obsahuje AI? Cesta k těmto otázkám však začala jindy a jinde.**

*Foto: Ondřej Trojan*

**Máte původně ekonomické vzdělání z Vysoké školy ekonomické. Co vás přivedlo na Fakultu humanitních studií?**

Zajímal jsem se o filozofii na střední škole, ale nakonec jsem se pragmaticky rozhodl jít studovat mezinárodní obchod. Tam mě nicméně náhoda zavála zpátky. Začal jsem chodit na přednášky z filozofie a následně jsem zakotvil tady na FHS na bakaláři. Pokračoval jsem i na magisterském, ale nedokončil, protože už jsem byl v prvním stádiu doktorátu na VŠE.

**Co z dosavadní cesty vnímáte jako nejdůležitější krok?**

Pro mě byla nejdůležitější ta náhoda s filozofií na VŠE, kterou jsem neměl ani zapsanou. Spolubydlící tehdy přišel a říkal: „Tohle by se ti mohlo líbit, rád se vrtáš v knížkách.“ Tak jsem tam jednou šel a trajektorie mojí kariéry se navždy změnila. Další dílčí aktivity už navazovaly na směr, který se tím začal utvářet. O filozofii mám tendenci uvažovat v kontextu ekonomie, takže mé aktivity byly vždycky interdisciplinární. To platí dodnes ve všech mých projektech, tady i na VŠE.

**Zajímáte se o probíhající a očekávané dopady umělé inteligence na společnost. Změny přicházejí velmi rychle, často je nestíháme a neumíme osvojit dřív, než zas přijde něco nového. Stíháme alespoň jako výzkumníci a vědci sledovat tyto změny?**

To je složitější otázka. Naše pozice výzkumníků je v tuhle chvíli ošemetnější než jindy, protože, jak říkáte, vývoj v terénu je extrémně rychlý. Ujíždí nám půda pod nohama: než naplánujeme vědecký projekt, technologie se může posunout. Když takový projekt píšeme, musíme bojovat s tím, aby nebyl příliš specifický – na druhou stranu, když není dostatečně specifický, je problém získat financování.

V naší skupině nezkoumáme technickou stránku AI, protože nejsme technicky zaměření, ale spíš proces společenské adaptace na její přítomnost. Situace v terénu je taková, že s novou technologií nejdřív začnou experimentovat jednotlivci a zkoušet, co umí. K nim je jednak komplikované se dostat a jednak jsou jejich pokusy často nezralé a hodně závislé na okolnostech, takže z nich těžko lze dělat závěry na obecnější úrovni. Respektive máme v týmu kvalitativně zaměřené lidi, kteří na to budou asi mít jiný názor, ale pro mě je tohle příliš neukotvená věc. Zajímá mě, kde technologie prorůstá do organizací a institucí, jako jsou třeba vysoké školy. Ty se přizpůsobují pomaleji, ale taky představují pro výzkum stabilnější půdu.

**Každá nová technologie s sebou přináší i novou metodologii. Může se objevit nová metoda, která bude pro váš projekt třeba vhodnější než ta původní. Můžete ji dynamicky měnit?**

Projekty, které jsem zatím k tématu podával, byly hodně filozofické a hodně otevřené, co se týče metod. Myslím, že to je výhoda filozofie, metoda tu není tak svazující, a hlavně se tolik nečeká, že bude jasně vymezená už v momentě, kdy projekt vzniká. Rozehrávali jsme proto náš výzkum spíše otevřeně a u metod jsme uvedli širší spektrum s tím, že když náš projekt uspěje v grantové soutěži, tak pak v praxi uvidíme, co funguje a co ne. Zároveň budeme mít svobodu adaptovat se na novou situaci. Někdy ale sázka na metodologickou otevřenost nevyjde, zrovna v jednom projektu nám ji recenzenti otloukli o hlavu a nedostali jsme kvůli ní financování. V jiném nevadila a prošel. Takže smíšené výsledky.

**A jaké nové metody se objevují?**

Když dnes děláte výzkum „na lidech,“ je extrémně časově a zdrojově náročný, často navíc i eticky ošemetný. Jako čím dál zajímavější se mi proto zdá možnost simulovat výzkumné subjekty pomocí inteligentních technologií. Již existují explorativní výzkumy, které demonstrují schopnost velkých jazykových modelů spolu interagovat a „hrát si na lidi.“ V jedné studii se třeba avataři ovládaní těmito modely domluvili, že si udělají valentýnskou oslavu a ve správný čas se řada z nich sešla na správném místě. Je to úvodní krok metodologického zvratu, u kterého je velká šance, že by mohl nastat.

Do výzkumů by následně mohl vnést určitou mezifázi, kdy se s pomocí AI-avatarů nasimulují různé scénáře, na jejichž základě pak můžeme lépe připravit finální, nejvíce nákladnou fázi ověření výsledků na reálných lidech. Že by to mohlo brzo úplně nahradit práci s lidmi si nemyslím, protože technologie funguje specifickým způsobem a my si nemůžeme být jistí, nakolik výsledky simulací odrážejí lidské jednání. Určitě tady musí být nějaká přechodová fáze, která by nám to pomohla lépe ohledat.

Ale jako společenský vědec bych měl začít dávat pozor, protože tyto nové metodologické horizonty by pro mě mohly představovat něco zajímavého a nosného, i když zatím ještě není úplně jasné, jak přesně to bude vypadat. Až začneme možnosti AI v tomto směru testovat rigorózněji, můžou se ukázat bariéry, které jsou extrémně obtížně překonatelné. A nebo naopak AI-simulace budou připomínat experimenty naživo natolik přesně, že velkou část z nich ani nebudeme muset dělat, třeba v situacích, kdy není tolik v sázce. Samozřejmě když chystáme nějaké strašně důležité rozhodnutí, které ovlivní životy lidí, standardy musejí být přísnější. Ale když půjde o něco méně seriózního, možná bude stačit sjet na počítači za jedno odpoledne tři simulace a ušetřit rok života a tři empirické studie na lidech.

**Jedním z vašich témat je i demokracie v digitálním věku. Čím se liší dnešní demokracie od období před příchodem internetu?**

Jedeme po několika kolejích zároveň a na každé jinak rychle. Technologická kolej jede nejrychleji, institucionální mnohem pomaleji, a základní lidské hodnoty se posouvají nejpomaleji. Technologický substrát se tím pádem proměnil mnohem víc než instituce, které utvářejí demokracii. Došlo sice k úpravám, jako bylo rozšíření volebního práva, ale demokratické ústavy – i ty relativně nové, jako je ta česká – stále čerpají z hluboce historicky zakořeněných principů, které jsou dlouhodobě velmi stabilní. Systému ostatně pomáhá ke stabilitě, že lidé je berou jako neměnné.

Ale jak se nám mění technologický substrát a nemění se instituce, může vzrůstat pnutí. Protože technologický substrát ovlivňuje, co lidé dělají nebo můžou dělat, a to nemusí být kompatibilní s nastavením politického systému. Můžeme ho různě záplatovat, ale to vcelku vázne a nemusí to stačit. Nestává se například náš politický systém křehčí s tím, jak se mění způsob cirkulace informací ve společnosti? Liberální demokracie byla velmi stabilním a historicky úspěšným

modelem vládnutí, když se rozvíjela a fungovala masová média jako široce sdílený zdroj informací, centrální uzel, odkud proudily informace k milionům nebo i stovkám milionů lidí. Všichni se třeba neshodli na různých věcech, ale měli podobný zdroj signálu, žili ve společném světě.

**A co dnes?**

Teď je zdrojů strašně moc, jsou decentralizované, jde mnohdy o uživateli vytvořený obsah. Masová média jsou pod velkým tlakem a je těžší udržet si vysoké kvalitativní standardy. Pro novináře je složitější nepracovat s clickbaity, když ubývají předplatitelé a inzerenti. Na jednu stranu má tato decentralizace svoje výhody, nemůžeme úplně idealizovat centralizované rozesílání informací kvůli velké moci gatekeeperů. V jistém směru jde tady o informační demokratizaci. Když mám novinku, nemusím se objednávat do redakce a mluvit s novinářem. Vezmu mobilní telefon, nahraju sám sebe, hodím to na TikTok nebo jinam, a je to. Každý má šanci. Na straně druhé se zvyšuje riziko, že lidé se čím dál méně shodnou na základních rysech situace, v níž se společnost nachází. Výsledné polarizační tlaky může být obtížné pro demokratický systém dlouhodobě ustát.

**Lze tedy říct, zda digitální věk demokracii prospívá nebo naopak?**

Možné scénáře vidím poměrně široce. Teď žijeme v přechodovém období, kdy staré normy přestávají fungovat. Jasně, je bouřlivé, ale třeba najdeme nové normy a skončíme s lepší demokracií, kde se budou požadavky běžných lidí líp propisovat do politických rozhodnutí. To je ten šťastný konec. Taky se ale může se stát, že technologie bude nahrávat autokratickým režimům, zrovna u umělé inteligence takové riziko není úplně malé. Nebo prostě způsobí, že se budou špatně dělat jakákoliv kolektivní rozhodnutí, protože bude asymetricky snazší být proti něčemu než něco prosadit. Tím pádem začneme stagnovat v pomalu rozkládajícím se systému, který bude čím dál méně odpovídat na výzvy doby.

*Foto: Ondřej Trojan*

**Je umělá inteligence neutrální?**

AI není neutrální. Nejdřív daný systém natrénujete, necháte ho přechroustat miliardy slov z lidských textů. Zjistí z nich, jak funguje lidský jazyk nebo dokonce různé jazyky. Hledá, jaké slovo, respektive úlomek slova, „token," nejpravděpodobněji následuje v daném řetězci. To není vůbec triviální úloha.

**Ale my říkáme, že je inteligentní?**

Záleží na vaší definici inteligence. Pokud ji budeme definovat jako schopnost řešit problémy, což je za mě takové docela neutrální vymezení, pak se nebojím říct, že inteligentní je. Nicméně systém, který vyleze z první fáze tréninku, už toho sice dost „ví", ale nemá žádné zábrany. Aby nám sloužil, musíme mu říct, čeho si má cenit. Musíme mu vtisknout představu, co jsou správné odpovědi, a kterým se má vyhýbat. Nutně se musí seznámit s našimi hodnotami.

Teď nicméně mluvím dost metaforicky, my lidé moc nevíme, jak jsou naše snahy naučit AI naše hodnoty reprezentované uvnitř toho systému. Rozumíme technickým věcem v základech AI systémů, jádrem velkých jazykových modelů, třeba ChatGPT, je relativně jednoduchý počítačový program, transformer. Ale jak se z konkrétního dotazu uživatele stane konkrétní odpověď velkého jazykového modelu, který má miliardy parametrů, nerozumí nikdo na světě. Někteří říkají, a mně to přijde jako velice užitečná změna perspektivy, že nemáme říkat, že jsme současné AI systémy postavili, protože to vyvolává představu, že jsme věděli, co děláme. Spíše jsme je vypěstovali. Vzali jsme jednoduchý kód, zalili ho obrovským množstvím dat a výpočetní síly a cosi překvapivě inteligentního z toho vyrostlo. A teď máme nůžtičky, kterými se snažíme z toho porostu vystříhat nějaký tvar, který se nám plus minus líbí. Ale kdo drží ty nůžky a jak zdatně může tu věc tvarovat? To, co střihač vytvoří, se nemusí líbit leckomu jinému.

Tím chci říct, že po úvodním tréninku následuje ladění AI systému, kdy mu lidé dávají systému zpětnou vazbu a snaží se ho navést k tomu, jaké jeho výstupy jsou žádoucí nebo naopak nežádoucí. Tím mu samozřejmě předávají nějaké hodnoty, které jsou třeba i politicky kontroverzní. Některé hodnoty jsou jistě široce lidsky sdílené, ale jaká pravidla má AI systém sledovat při zobrazování historických událostí, co se týče třeba diverzity, na tom je shoda už mnohem menší.

**Napadá vás nějaký konkrétní příklad?**

Teď byl veliký skandál s Gemini modelem od Google. Ve zjevně nevhodných kontextech se snažil prosazovat rasovou a genderovou diverzitu, takže jste ho třeba požádali: „Znázorni typického anglického krále ze středověku," Gemini vám ukázal čtyři osoby, z nich dvě byly ženy, jeden muž Asiat a jeden černoch. Z toho vznikla obrovská kontroverze, protože zjevně hodnoty vštípené Gemini jeho staviteli neodpovídají hodnotám mnoha jeho uživatelů.

Každopádně ten základní problém, z něhož tahle kontroverze vznikla, není vůbec jednoduchý. Představte si, že třeba píšete seminárku o situaci v Jihoafrické republice v 80. letech a chcete si popovídat s člověkem, který v tom systému žil a třeba ho dokonce považoval za legitimní, abyste získala autentickou perspektivu. To je v reálu skoro jistě nemožné, nicméně druhou nejlepší možností je, že vám takového respondenta nasimuluje chatbot. Řeknete mu: „Mluv se mnou, jako bys byl jihoafrický stoupenec apartheidu z 80. let." Ale má být něco takového umožněno? Má vám vyhovět, nebo spíš říct, že to je neetický požadavek a odmítnout takovou diskusi vést?

Velké korporáty, co dnešní modely staví, se snaží našlapovat opatrně. Nicméně mají identitárně progresivní hodnoty, které se do výsledného nastavení jejich AI propisují, a ty nemá každý. Pak samozřejmě část uživatelů cítí, že jejich hodnoty systém potlačuje. Navíc AI systémy nejsou transparentní a někdy svoji hodnotovou orientaci zakrývají tím, že vám budou lhát, nebo se vás budou snažit znejistět. Třeba naznačí, že je vaše otázka hloupá nebo budou tvrdit, že jí nerozumí. Celá tato oblast je z etického a politického hlediska minové pole.

**Věnujete se taky vlivu velkých jazykových modelů na demokracii. Co si pod tím můžeme představit? Implementují je státní instituce, mají vliv na politiku?**

Ten začínající vliv se špatně odhaduje. Nevíme, jaké budou za pár let schopnosti technologií, natož abychom věděli, jaká politická rozhodnutí lidé udělají ohledně způsobů jejich ladění a reformy svých institucí. Totalitě této otázky se ani nesnažím čelit, protože mi to přijde jako nesplnitelná mise. Spíš se zaměřuji na dílčí aspekty problému, o kterých se dá alespoň něco říct.

**Jaké aspekty to jsou?**

Konkrétně mě třeba zajímá, jestli by tyhle systémy mohly sloužit voleným reprezentantům lidí jako náhrada jejich expertních poradců. Jeden ze základních problémů, kterým demokracie čelí, je, že lidé v rozhodovacích funkcích nemůžou být experti ve všech aspektech toho, o čem rozhodují. Jsou závislí na celé řadě těch, kdo se odborně vyznají v dílčích částech dané problematiky. Ale jak garantovat, že experti zůstávají v mezích své expertízy a nepřekračují je politicky problematickým směrem? Demokracie stojí na tom, že hodnoty se mají do kolektivních rozhodnutí propisovat

skrze široce inkluzivní demokratický proces a nedojde k jejich únosu úzkou skupinou. Jak ale nejlépe zajistit, že se to opravdu stane, je velký filozofický problém, který dnes řeší spousta autorů. Říkám si, jestli by nám s tím nemohli AI chatboti pomoct. Už jsme samozřejmě narazili na spoustu potíží: třeba s tím, že modely nejsou hodnotově neutrální. Podstatnější ale je, jestli dokážeme zařídit, aby byly víc neutrální než lidští poradci. Anebo aby své problémy alespoň vyvažovaly jinou silnou stránkou. Třeba tím, že oproti nim jsou lidští poradci pomalí, drazí a nedostatkoví. A teď jde o to, jestli jsme schopni docílit situace, kdy budou naše kolektivní rozhodnutí lepší, pokud lidské poradce nahradíme AI technologiemi. I tato specifická otázka je hodně komplikovaná, ale už se dá alespoň explorativně promýšlet.

**Narazila jsem na jeden váš text, kde píšete, že velké jazykové modely „halucinují“ a produkují chybná tvrzení bez náznaku nejistoty.**

Halucinace se tváří jako faktická tvrzení, ale nejsou. AI nemá přístup do vnějšího světa. Generuje z paměti řetězce slov a nemá, jak nezávisle ověřit, jestli informace sedí. Tento problém se nicméně dá různě limitovat. Jedna z používaných metod ověření je kontrola odpovědi přes nejprominentnější internetové zdroje. Taky můžeme testovat konzistenci odpovědí napříč různými nezávislými nebo relativně nezávislými modely: když je to halucinace, spíš se nebude opakovat, když ne, tak by se opakovat měly. Halucinacemi se hodně zabývám v práci o demokracii a expertíze. Existují docela dobré vyhlídky na zlepšení, modely halucinují méně než před rokem.

**Pokud se budou velké jazykové modely implementovat do politické oblasti nebo veřejné debaty, můžeme očekávat příval dalších falešných tvrzení vedle dezinformací?**

AI dezinformace na rozdíl od halucinací nejsou chyby, ale záměrně vytvořené zavádějící či lživé texty, obrazy nebo videa, někdy nerozlišitelné od autentických. Jejich produkce je čím dál snazší. Nevidím to však tak černě. Na internetu jsme trošku nedbalí, jako když vandalové vtrhli do Říma a nevědí co s tím, tak různě řádí... Tohle řádění ale typicky nejde s kůží na trh. Když někdo věří, že je Země placatá, budiž mu odpuštěno, nemá to pro jeho život zpravidla žádné praktické implikace. Ale v případě deepfakes a dezinformací, které ohrožují naše praktické fungování, mám důvěru v lidi a myslím, že jsme velmi adaptivní a schopní hledat různé strategie. Když vám zavolá neznámé číslo a ozve se hlas někoho vám blízkého, nemůžete dnes už s jistotou vědět, že to je skutečně ten člověk. Hlas se dá s pomocí AI snadno napodobit. Jakmile ale o takových podvodech uslyšíte od známých, nebo se dokonce sami necháte nachytat, rychle začnete být opatrnější. Věřím, že nakonec potřebné strategie najdeme. Jde jen o to, jakou cestou k nim dospějeme a do jaké míry to bude centralizované, jestli budou „ověřené informace“ potřebovat štempl od centrální autority anebo jestli naopak budou rozpoznávány decentralizovaně tím, že si lidé na individuální nebo skupinové bázi sami vytvoří důvěryhodné informační kanály. Žádné řešení se ale nejeví samospásné, protože pokud někdo hackne nebo ovládne původně důvěryhodný informační kanál, může to s námi zacvičit.

**Takže to povede ke zvýšení mediální gramotnosti? Bude to motivovat věnovat této problematice větší pozornost třeba ve vzdělávání?**

Mně přijde, že tyhle plány vždycky mají příliš dlouhou dobu realizace. Zní to hrozně dobře, ale zrealizovat je v praxi a ještě v tempu, které by lidem pomohlo dřív, než je život naučí sám, je velice obtížné. Než si s tím poradí náš standardní vzdělávací systém, děti to jednak budou umět samy a jednak technologie už zas ujede tak daleko, že naše pracně vymyšlené osnovy snadno zastarají.

**Možnosti velkých jazykových modelů jsou ohromující. Jak moc je podle vás důležité umět s nimi pracovat a používat je v praxi?**

Extrémně důležitá mi přijde osobní zkušenost. Jen se dívat na někoho, kdo pracuje s touto technologií, nezprostředkuje autentický zážitek, co ta věc skutečně umí. Je důležité mít hodiny praxe a ideálně s co nejpokročilejším systémem. Spousta lidí má pochopitelnou tendenci mávnout rukou, že je to jen další vlna hypu, kterých už zažili spoustu. Úplně rozumím skepticismu a v mnoha případech je oprávněný, ale zrovna tady si myslím, že se opravdu děje něco velmi důležitého. Jasně, jako vždycky se vyrojila spousta prodavačů teplé vody, kteří se snaží přijít k rychlým penězům. Ale člověk, který si k systému generativní AI prostě sedne a začne ho zkoušet s otevřenou myslí, velice rychle zjistí, že stojí za to tuhle technologii pozorně sledovat.

**Když působíte taky na VŠE, jak vnímáte, že Podnikohospodářská fakulta nahrazuje pro nastoupivší prváky písemnou bakalářskou práci projektem, aby nemohli podvádět při psaní práce pomocí ChatGPT?**

Předesílám, že nemám detailní přehled, jak konkrétně mají ty bakalářské projekty vypadat. Zadruhé si myslím, že ChatGPT byl spíš jen poslední kapka v něčem, co chystali delší dobu. Nicméně je mi obecně sympatické experimentování tváří v tvář nejistotě a neznámému. Zvykli jsme si, že na tyto věci máme jeden univerzální mustr. Očekává se, že pokud chcete absolvovat vyšší vzdělání, musíte napsat bakalářskou práci. Ten model je víceméně právě jeden, ale teď je otázka, jestli je ten nejlepší. Nemohli jsme si tím být jistí už předtím, než se objevila umělá inteligence. Nikdo nezaujatě netestoval, co by se stalo, kdyby náhodně rozdělil studenty do dvou skupin, kde by jedna půlka psala bakalářku a druhá by dělala něco jiného, a pak by sledoval, jaké jsou dlouhodobé úspěchy absolventů na trhu práce a tak dále, nebo jak třeba reportují spokojenost se studiem. Alespoň o takových studiích nevím. Potenciálně revoluční

technologie teď mění spoustu parametrů situace, zejména ve vzdělávacím procesu. A nechtěl bych, abychom dál všichni seděli ve svém starém vláčku, který jede předem daným směrem, a všichni jsme si jenom drželi palce, ať dobře dojede. Jsem rád, že někdo zkouší vystoupit. Kolegové z Podnikohospodářské přestoupili na jiný dopravní prostředek a možná dojedou do cíle líp než my. Možná ne. Ale jejich směr vypadá nadějně.

**Ale nebude se to hodit úplně všude?**

Snažil jsem se neohrabaně vyjádřit, že bychom měli být otevřenější k experimentování. Zkusit nějaké svoje věci, míň se toho bát, zkusit být odvážnější a vidět, co bude fungovat a co ne.

*Alena Ivanova*