Výzkumy ohrožující podstatu života člověka, tedy zásahy do genetického kódu, by se neměly povolit

Vladimír Mařík Díl 1/3

Text 16.7.202431 min Přehrát
#Společnost

Přichází éra umělé inteligence, na tom se shodnou asi všichni, laici i odborníci. Tím ovšem ale shoda pro mnoho lidí končí. Jedni tvrdí, že výhody umělé inteligence budou daleko, daleko přesahovat případná marginální negativa, jiní jsou opatrnější, a vidí jak plusy, tak mínusy nového věku vcelku vyrovnaně. Ale nemálo lidí se umělé inteligence zkrátka bojí s tím, že žádné pozitivum, jakkoli se to dnes tak jeví, nemůže vyvážit vážná rizika pro náš dosavadní život, a mnozí skeptici mají na mysli dokonce samotnou existenci naší civilizace. Tak jak to tedy je? Máme vítat umělou inteligenci s otevřenou náručí, nebo ji hnát od sebe a bránit se jí? Mým hostem je pan profesor Vladimír Mařík, vědec, zakladatel a vědecký ředitel Českého institutu informatiky, robotiky a kybernetiky, zkráceně CIIIRC, Českého vysokého učení technického v Praze. Umělou inteligencí se už zabývá dlouhou řadou let, stejně jako výzkumem umělé inteligence a vědomí. Podílel se na celé řadě knih na toto téma, například „Eseje o vědomí směrem k umělé inteligenci“, nebo v nejnovější knize, která zatím ještě nespatřila světlo světa, ale už se to blíží: „Proč se nebát umělé inteligence“.

Martina: Pane profesore, četla jsem synopse vaší nové knihy, vlastně její hrubou verzi, a vy tam s dalšími autory píšete, že můžeme umělou inteligenci vnímat čtyřmi různými způsoby: Jako efektivního pomocníka, který poskytuje užitečné nástroje. Jako zdroj výzev pro

rozvoj poznání a společnosti. Můžeme ji hodnotit jako hrozbu. Nebo ji také posuzovat jako zdroj nereálných očekávání. Řekněte mi, ke kterému z těchto chápání umělé inteligence se vy teď přikláníte nejvíc?

Vladimír Mařík: Samozřejmě k tomu prvnímu, protože umělá inteligence se ukazuje jako pomocník. Ale v tom druhém bodě jsme říkali, že také je to zdrojem výzev, a zdroj výzev je také velmi důležitý, protože nás posílá zkoumat věci, které jsou pro lidi zajímavé, pro rozvoj umělé inteligence podstatné, a současně tento výzkum umožňuje naznačit, proč se umělé inteligence nemáme bát, nebo proč se jí bát musíme. Jsou tu samozřejmě některé aspekty, které hovoří o nebezpečí umělé inteligence, ale my můžeme jasně identifikovat, kde toto nebezpečí je, v čem spočívá, a můžeme mu krásně čelit. Samozřejmě výzkum, který vychází z těchto výzev, může také rozptylovat vědeckofantastické představy o tom, jak inteligence jednou změní celý svět, stane se svébytnou entitou, která začne ovládat planetu Zemi a celý vesmír, a jak se bude tato superinteligence, jak se tomu říká, dál šířit.

Martina: A tomu vy nevěříte?

Vladimír Mařík: Tomu tehdy rozhodně nevěřím. Dokonce jsme se v té knize mezi autory shodli v tom, že to je svým způsobem jenom fantazie, která není založena na vědeckých poznatcích – je to svým způsobem pavěda.

To, co do AI vložíme – data, zkušenosti, znalosti – systém AI využije k simulaci inteligentního chování, ale sám nikdy svou rozhodovací činnost neprožívá, a neuvědomuje si ji

---

Martina: „Je to svým způsobem pavěda.“ Dobře, ale řekněte mi, mohl byste vy ve své pozici, v okamžiku, kdy tady budujete tento institut, říct vůbec něco jiného?

Vladimír Mařík: Já se domnívám, že ve své pozici říkám to, co má vědecky podložené základy. Prostě musíme se držet toho, co nám říká současná věda, a i kdyby říkala, že tady jednou kosmická inteligence vznikne nějak sama, tak dokud pro to nejsou jakékoli podklady, tak to nemůžu podporovat.

Martina: Toto samozřejmě předznamenává, v jakém duchu se náš rozhovor ponese. Já se budu ptát vědce, který je přesvědčen o tom, že nám umělá inteligence může být dobrým pomocníkem a dobrým nástrojem. A samozřejmě se také budu pokoušet zjistit, jestli opravdu na všechno máme odpovědi, které lze změřit, zvážit, sečíst, a nějakým způsobem vědecky popsat. Protože na začátku vaší poslední knihy, kterou jsem tady zmiňovala, je citát, jehož autorem je Michael Hsieh ze Stanfordské univerzity, a teď ho cituji: „Jednou ze znepokojujících vlastností umělé inteligence je, že halucinuje fakta. Umělá inteligence ve skutečnosti ani nezdůvodňuje, ani nepřemýšlí, jen dokáže velmi dobře imitovat lidské způsoby komunikace. Dělá na nás dojem úžasnými způsoby popisu a komunikace, ale stále jen napodobuje.“ To je to, co si vy myslíte? Tedy že umělá inteligence stále dělá jenom to, co do ní vložíme?

Vladimír Mařík: Ano, určitě, tento citát je pravdivý. To, co vložíme, data, která vložíme, naši zkušenost, naše znalosti, systém umělé inteligence nějakým způsobem využije k simulaci svého chování, k simulaci inteligentního chování, ale sám systém nikdy svou rozhodovací činnost neprožívá, a neuvědomuje si ji, a to, co do něj vložíme, se nějakým způsobem odrazí. Protože systémy – a tady měl pan Hsieh asi na mysli systémy GPT, a velké jazykové modely, a velké modely obecně – jsou natrénované. Ale my nevidíme dovnitř tohoto systému, není tam žádná možnost vysvětlit nastavení parametrů. Čili, systém GPT často poskytne odpověď, která je zdánlivě rozumná, ale ve skutečnosti nedává žádný smysl, nebo je dokonce nepravdivá.

A protože odpovídá tomu modelu, který si sám stroj natrénoval, tak říkáme, že halucinuje. Že tedy vykládá něco, co není pravdivé, ač to pravdivé vypadá, ač to tento systém za pravdivé považuje.

V natrénované neuronové síti GPT nepřečteme vůbec nic, nedovedeme vysvětlit žádné parametry, nerozumíme tomu, jak dospívá k rozhodnutí. A tomuto nerozumění se snažíme bránit.

Martina: Já si vzpomínám, že když jsem si tady povídala s odborníky o DNA, tak vlastně oni považují za přežitelné z DNA jenom čtyři procenta, která jsme dokázali rozkrýt, ten genetický kód. A zbytek považují za takzvaný DNA trash, a to mi vlastně přijde možná přemoudřelé, protože když my 96 procentům nerozumíme, tak to označíme za smetí. Neobáváte se, že se můžeme do podobné situace dostat právě i u umělé inteligence?

Vladimír Mařík: Tam jsme na tom hůř. Tam nepřečteme na trénované neuronové síti vůbec nic, tam je nula. Čili trash by mělo být 100 procent, ale my si toho přesto vážíme. My ten natrénovaný model nedovedeme vysvětlit, nedovedeme vysvětlit žádný z parametrů, které v neuronové síti jsou, ale to neznamená, že nemají žádnou hodnotu, že nemají žádnou roli při finálním rozhodování. Takže my se snažíme bránit se tomu, že nerozumíme tomu, jak GPT dospívá k rozhodnutí. A bráníme se tak, že se snažíme budovat znalostní grafy, nebo strukturované znalosti, které kontrolují, zda výsledek GPT dává smysl. To je například u nejnovějšího systému, Claude, který už je o něco lepší než třeba GPT 4. kontroluje smysluplnost vygenerovaného textu z hlediska struktury obecných znalostí. A tyto obecné znalosti jsou v takové podobě, že jim člověk rozumí, jsou vysvětlitelné, jsou uchopitelné. Čili struktura uchopitelných, srozumitelných znalostí posuzuje, zdali model, do kterého nevidíme, nemluví úplně z cesty, nehalucinuje, ale dává to smysl, je to konzistentní s naší znalostí o světě.

Martina: Já našim posluchačům slibuji, že se budeme bavit i o konkrétních věcech, o dopadech umělé inteligence, kterou už máme k dispozici a kterou budeme využívat, a to ať už na práci, případnou nezaměstnanost, ale také na vznik určitého informačního chaosu, a tak dále. Tomu všemu se ještě budu věnovat. Ale ještě bych přece jenom zůstala možná u teoretických věcí, protože v odborných kruzích – to jsem se také dočetla ve vaší knize – se říká, že dojde k takzvané Kurzweilově singularitě, což je unikátní zlomová situace ve vývoji technologií, jejíž součástí bude i to, že stroje předčí lidskou inteligenci. Je pravda, že tento vědec a futurista Kurzweil odhadoval, že k tomu dojde cirka v roce 2045 – ale už to trochu posunul. Přesto všechno je to velmi zajímavá úvaha, a on ji považuje za možnou, to znamená, že stroj předčí lidskou inteligenci. A teď se na mě díváte tak, že si to nemyslíte, a že se mýlím já, i Kurzweil.

Vladimír Mařík: Tak především Kurzweil predikoval v roce 2005, že v roce 2045 tady budou stroje, které budou inteligentnější než člověk. A dokonce říkal, že ho můžou začít ovládat, a mít nad ním nadvládu.

Martina: Takto už jsme si představovali rok 2000, jak budeme mít vznášedla místo osobních vozů, a všecko kolem nás budou obsluhovat roboti.

Vladimír Mařík: Ale on pak couval. O deset později začal couvat, že to nebude v roce 2045, ale že to bude tak ke konci tohoto století. A myslím, že to bylo předloni, kdy couvnul ještě víc, a říkal, že v tomto století se toho nedočkáme, a nebudou to systémy čistě anorganické na bázi křemíku, musí to být systémy, ve kterých se bude kombinovat živá hmota a její aktivity s křemíkem. Čili, začíná posouvat svůj názor směrem k biologicky propojeným organismům s počítačem budoucnosti, čili jde mu o symbiózu živé a neživé hmoty. Tvrdí, že v tomto století to ještě nebude, ale pak možná – není si úplně jist – by to mohlo nastat.

On svým způsobem rozpoutal výzkum v tom, jaké jsou podmínky toho, aby stroj začal ovládat člověka. V tomto smyslu považuji jeho výzkum za přínosný, motivující, je to výzva, o které jsme hovořili, kterou umělá inteligence přináší. A tento výzkum začal probíhat v nejrůznějších disciplínách, nejenom u umělé inteligence, ale začal tam být výzkum například i v biologických vědách, lékařských vědách, snaha o to, stanovit si, kde jsou hranice neživé hmoty a živé hmoty, kde začíná život, kde se neživá hmota mění na živou. Zajímavé při tom je, že do toho zasáhl i český ekonom Milan Zelený, který ve své knize napsal o tom, jak si představuje, že z anorganické hmoty vznikl život čili hmota organická. Že to bylo v uzavřené bublině pod nějakým tlakem, s nějakou propustností této bubliny, a je to hodně citovaná česká stopa právě v otázce, kde vznikl život. A to nám umožňuje studovat, kde je rozhraní mezi živou, neživou hmotou.

Existují výzkumy, které mohou ohrozit podstatu života člověka, tedy zásahy do genetického kódu, metody CRISPR, umožňující vyměňovat subřetězce a vytvářet ideálního jedince

---

Martina: A tomu se do budoucna budeme muset asi hodně věnovat, protože tam vznikne, řekla bych, mnoho etických a nebezpečných možností, věcí, které by se nám možná mohly vydat směrem, který by mohl být pro lidstvo nebezpečný. Řekla jsem to správně, nebo to byla jenom taková slovní vsuvka? Mohou nás v tomto výzkumu potkat hraniční předěly? Hranice, kdy se budou muset vědci rozhodnout, jestli je ještě možné tuto hranici překonat, nebo by mohli strhnout lavinu, kterou pak nebudou moct třeba už ani ovládat, a už vůbec ne zastavit?

Vladimír Mařík: Samozřejmě, že máte pravdu, akorát si myslím, že není třeba hovořit o budoucím čase, protože už dnes existují výzkumy, které mohou ohrozit podstatu života člověka, což jsou třeba zásahy do genetického kódu. Už jsou metody typu CRISPR, které umožňují vyměňovat subřetězce a připravovat ideálního jedince, a to jsou

výzkumy, které ale neměly být povolovány, nebo bychom se z nich neměli radovat. Pokud jsou to výzkumy v laboratoři, prosím, ale nemělo by se to později dál aplikovat.

Martina: Nemělo by, ale nevíme, co se může odehrávat v nejrůznějších soukromých laboratořích na nejrůznějších kontinentech. Řekněte mi, chápete iniciativu Elona Muska, který se také stal spolusignatářem výzvy, aby se práce na umělé inteligenci, řekněme, zbrzdila, zastavila, pozastavila, protože by se lidstvo mohlo dostat na scestí? Je to populismus, nebo je to obava člověka, který ví?

Vladimír Mařík: Já se domnívám, že to je do určité míry populismus, a do určité míry obchodní tah, protože Musk potřeboval získat čas, aby dohnal ty, co mu o půl roku, rok, rok a půl, utekli, a jsou před ním. Čili, má to několik dimenzí. Já samozřejmě nevidím do Muskovy hlavy, a nevím tedy, který z těchto argumentů je důležitější, nicméně si myslím, že obava o osud lidstva byla spíše tím slabším argumentem.

Martina: A vy ji máte, pane profesore? Jsou okamžiky, kdy pochybujete, zda je tato cesta správná, a zda lze pokrokem, poznáním vysvětlit a omluvit i tuto překotnou cestu?

Vladimír Mařík: Já se domnívám, že každý, kdo pracuje v této oblasti, by si měl uvědomovat, kde jsou hranice toho, kam ještě může jít, kam může jít věda, kam může jít se svým výzkumem, měl by nad tím přemýšlet. Ale domnívám se, že výzkum v umělé inteligenci nelze zastavit, ale musí se zastavit tam, kde začíná být evidentně nebezpečný, nebo kde si nejsme jisti, že by mohl být opravdu nebezpečný. A proto Muskova výzva měla směřovat k zastavení nebezpečného, nebo pro lidstvo zničujícího výzkumu, nikoli směrem k výzkumu, který přináší užitek. A jak jistě víte, minulý měsíc Evropská unie přijala akt pro umělou inteligenci, ve kterém klasifikuje systémy umělé inteligence podle rizika pro společnost, protože některé výzkumy

jsou prostě v nejrizikovější kategorii, a tam je všem evropským státům doporučeno, aby takový výzkum zakázaly, a dokonce právně upravily tak, aby nebylo možné ho provádět.

Martina: Tato členění, tyto skupiny jsou čtyři.

Vladimír Mařík: Od neškodných, kde o nic nejde, přes mírně nebezpečné, kde se požaduje splnění určitých pravidel, kritické, kde se požaduje testování, ověřování, dokumentace, a tak dále – a má to smysl. Lidé se nás ptají, jestli nás evropská byrokracie neomezuje, a já bych řekl, že to zatím žádný dopad na výzkum, který je pro člověka užitečný, nebo, který je smysluplný, který pracuje i s velkými daty, a může být v kritické kategorii, to zatím nemá, nijak nás to neomezuje, a myslím, že většina výzkumníků to považuje za správné.

Martina: Pane profesore Vladimíre Maříku, vy jste řekl, že dvě jsou víceméně neškodné…

Vladimír Mařík: Neškodná, a málo škodlivá.

EU zařadila do kritické kategorie výzkum AI vše, kde může dojít k poškození člověka, nebo lidstva

---

Martina: Ano, a málo škodlivé kategorie. A pak jste řekl „kritické“. Můžete – než se dostaneme k těm nepřijatelným, což asi budou zásahy do DNA, ale nechci vám napovídat – říct, které to jsou?

Vladimír Mařík: Kritické jsou především ty, kde se jedná o metody léčby a terapie, kde tedy může dojít k poškození zdraví člověka. Kritické jsou třeba tam, kde je hromadná doprava a hrozí, řekl bych, dopravní katastrofa většího rozsahu, čili všude tam, kde chybné rozhodnutí může vést k většímu, či menšímu poškození lidí, člověka, velkého počtu lidí, a podobně. Čili, do kritické kategorie spadají ty, které by mohly poškodit člověka, nebo lidstvo.

Martina: Už jsme tady zmínili nejrůznější světadíly, nejrůznější země, a bůhví jaké laboratoře a průzkumy, výzkumy, které tam mohou probíhat. Každý, kdo něco podobného vyvíjí, by asi měl být schopen převzít za to zodpovědnost. Je to vůbec možné? Ptám se proto, že je to dlouhý proces, na kterém se bude podílet celá řada lidí, celá řada institucí. Je v takovém případě možné, aby za to někdo nakonec převzal zodpovědnost?

Vladimír Mařík: To je velmi zajímavá otázka, kterou řeší právníci v mnoha zemích – kdo vlastně nakonec nese zodpovědnost za systém s umělou inteligencí, který byl někde vyvinut, někde byl převeden do masového použití, někdo ho provozuje, někdo ho bezprostředně spustil. Kdo je zodpovědný? Vynálezce, programátor, vlastník, uživatel, provozovatel? Kdo je zodpovědný?

Martina: Pane profesore, spadne nám most na koleje, a dodnes za to, myslím, nebyl učiněn prakticky nikdo zodpovědným, kromě jednoho – snad – viníka, který byl jenom kolečkem v soukolí. Desítky let stavíme dálnici, které se teď říká „moravské“, nebo „ostravské moře“, protože se prostě propadla, zvlnila, a viníka nikde nemáme, ani u takovýchto, řekla bych, banalit. Co teprve u takovéhoto výzkumu? Myslíte, že je to vůbec možné, nebo ve výsledku zase – v případě různých potíží – opět nikdo za nic nebude moct, nebo to bude představovat světelné roky soudních tahanic s nejrůznějšími novými kyberodborníky a kyberprávníky?

Vladimír Mařík: Myslím, že vaše představa o tom, že to bude složitý právní proces a že to nebude jednoduché, je správná – a bude možná v některých případech ještě horší než s dálnicemi. Nicméně, jedna z možností, jak přeci jenom někomu zodpovědnost přičlenit, je certifikace, nezávislá certifikace těchto systémů, což bude v první řadě důležité pro lékaře, kdy by za systém, který radí při nějaké diagnóze, nebo operaci, měl někdo nést konkrétní odpovědnost, čili měl by být

certifikován pod určitým jménem. Totéž bude platit u systémů, které budou řídit například jaderné reaktory – tam musí být certifikace, odpovědnost, jasně stanovena. A samozřejmě mezi tím budou vznikat tisíce dalších systémů, které budou na koleně vyrábět malé start-upy, které budou třeba stanovovat diagnózu, nebo dietu, nebo cokoliv jiného, ale tam už odpovědnost nebude vymezena tak ostře. A tam můžou nastat značné problémy.

Martina: Pane profesore Maříku, jestli tomu správně rozumím, tak vy říkáte, že už teď se na této certifikaci pracuje a že už teď Evropská unie vydala akt umělé inteligence, to znamená, že se snaží nové zákony pro činnost robotů svým způsoben definovat už dnes. Ale jestli tomu rozumím správně, tak asi tyto robotické zákony budou poněkud jiné než ty, o kterých psal Issac Asimov?

Vladimír Mařík: Budou jiné. Já bych chtěl jenom říct, že Evropská unie začala tento systém budovat. Ještě není žádné certifikační pracoviště, to se teprve musí v Evropě podle návodu vybudovat. A samozřejmě Asimova pravidla tam budou, ale bude to jen základ, bude jich tam asi mnohem víc, budou mnohem detailnější, a testování a verifikace budou náročné procesy.

Martina: Víte, co by mě zajímalo? Jestli na nových zákonech pro činnost robotů bude kooperovat umělá inteligence?

Vladimír Mařík: Tedy, jestli budou psány pomocí GPT? Tak to myslíte?

Martina: To by bylo asi jenom to nejmenší.

V Číně funguje sociální systém neustálého sledování, srážení bodů, a trestání za nežádoucí chování, včetně kritiky režimu. Toto bychom si v civilizovaném světě nepřáli.

Vladimír Mařík: Ano. Myslím, že tam bude potřeba použít zejména přirozenou inteligenci, a na podporu trochu té umělé. Ale bude hodně záležet na tom, jak to lidé rozmyslí.

Martina: Jelikož vím, že vy jste člověk přemýšlivý, a nejste jenom technik, ale také vlastně i ve vaší poslední knize, která vyšla, „Eseje o vědomí“ je vidět, že si kladete i otázky týkající se vědomí, technického vědomí, lidského vědomí, definice vědomí, a mohla bych ještě pokračovat. Řekněte mi, k čemu jste vy sám osobně došel? Jak by měly vypadat zákony, které upravují lidské nakládání s umělou inteligencí, aby společnost byla lépe chráněná? Je to vlastně vůbec možné? Nebo pojedeme metodou pokus – omyl?

Vladimír Mařík: Myslím, že se tyto zákony budou rodit postupně. Rozhodně dneska nelze říct, že máme všechno vyzkoumáno, a víme, jak mají být tyto zákony postaveny. Dneska víme některé nepřekročitelné pravdy, nebo pravidla, která nebudeme chtít v žádném případě překročit, ale ony se ještě vynoří drobné nuance, které mohou být velmi nebezpečné, a to nelze bez prvotních základů práva dále domýšlet. Čili, samozřejmě, na něco jsme přišli, ale rozhodně si nemohu dovolit tvrdit, že už máme představu toho, jak by toto právo mělo kompletně vypadat.

Martina: Přičemž u čtvrté kategorie, tedy „nepřijatelné“, se vypráví – „se vypráví“ není zrovna novinářsky podložený obrat – že tato pracoviště, která různě upravují, vstupují do DNA, nebo se o to alespoň pokoušejí, už existují. Myslíte si, že to tak je?

Vladimír Mařík: Je to tak. Samozřejmě, že to tak je. Důležité je, aby to byl výzkum pod kontrolou, aby to nebyl výzkum na třeba genech člověka, ale pokusných zvířat pod přísnou, velmi přísnou kontrolou. Do jaké míry to lze ochránit – nevím. Ale pokud se týká třeba sociálního sledování lidí, tak v Číně jsou tyto systémy vyvinuty, fungují. My jsme v podstatě pro výstrahu do té publikace zařadili jednu kapitolu o tom, jak to v Číně vypadá, a jak bychom si nepřáli, aby to v civilizovaném světě vypadalo: Kdy kamery sledují člověka od rána až do večera, a vědí o každém jeho pohybu, o tom, kdy přešel ulici na červenou,

srážejí mu body. A tito lidé, kteří mají nízké sociální ohodnocení, za trest nesmějí kupovat jízdenky do rychlovlaků, a mají jiná omezení, a to nemluvím o těch, kteří se nějak vyjádřili proti režimu.

Martina: A vy si myslíte, že za tímto nemáme nakročeno? Myslím tím teď Evropu.

Vladimír Mařík: Těžko říct. Rozhodně Čína je v tom tak daleko vepředu, že nám jako odstrašující případ ukazuje, kudy bychom jít neměli. Samozřejmě i Evropě jsou leckde snahy aspoň kousek tímto směrem postoupit. Ale Evropská unie tomu tím aktem řekla, domnívám se, naprosto jasné „ne“. Čili, bude to nelegitimní i na úrovni států.

Martina: Tak doufejme. Protože, jak jste řekl, Čína je už tak daleko, že nám ukázala, kudy za ní nekráčet. Ale obávám se, že být jakýmkoliv totalitním vládcem, tak už mám dárek k Vánocům vybraný. Je to inspirativní.

Vladimír Mařík: Proto se musíme snažit, aby v Evropě nebyli totalitní vládci. Snad nám i Evropská unie k tomu dá nějakou pojistku.

Martina: Amen.

Vladimír Mařík: Amen.

Všechny příspěvky s Vladimír Mařík

Diskuze:

Napsat komentář

E-mailová adresa nebude publikována.