FILOZOFIA
Volume 80, Number 4, 2025

DOI: https://doi.org/10.31577/filozofia.2025.80.4.1

Mohol by byť veľký jazykový model (LLM) vedomý?

DAVID J. CHALMERS, Department of Philosophy, New York University, New York, USA

CHALMERS, D. J.: Could a Large Language Model be Conscious?
FILOZOFIA, 80, 2025, No 4, pp. 459 – 481

There has recently been widespread discussion of whether large language models such as the GPT systems might be sentient or conscious. Should we take this idea seriously? I will discuss the underlying issue and will break down the strongest reasons for and against.

Keywords: consciousness – artificial intelligence – Turing test

Keď v júni 2022 Blake Lemoine, softvérový inžinier zo spoločnosti Google, vyhlásil, že v systéme LaMDA 2, jazykovom modeli založenom na umelej neurónovej sieti, zistil vnímanie a vedomie, jeho tvrdenie sa stretlo so všeobecnou nedôverou. Hovorca spoločnosti Google k tomu uviedol:

> Náš tím preskúmal Blakeove obavy a informoval ho, že dôkazy nepotvrdzujú jeho tvrdenia. Bolo mu povedané, že neexistujú žiadne dôkazy o tom, že LaMDA je vnímavá (a že existuje veľa dôkazov, ktoré to vyvracajú) (*Washington Post*, 11. júna 2022).

Otázka dôkazov vzbudila moju zvedavosť. Aké sú alebo by mohli byť dôkazy v prospech vedomia vo veľkom jazykovom modeli a aké by mohli byť dôkazy proti nemu? Práve o tom tu budem hovoriť.

Jazykové modely sú systémy, ktoré priraďujú pravdepodobnosti sekvenciám textu. Keď dostanú nejaký počiatočný text, použijú tieto pravdepodobnosti na generovanie nového textu. Veľké jazykové modely (LLMs), ako napríklad známe systémy GPT, sú jazykové modely využívajúce obrovské umelé neurónové siete. Ide o obrovské siete vzájomne prepojených jednotiek podobných neurónom, ktoré sú vytrénované na základe obrovského množstva textových údajov, a ktoré spracúvajú textové vstupy a reagujú textovými výstupmi. Tieto systémy sa používajú na generovanie textu, ktorý sa čoraz viac podobá ľudskému. Mnohí tvrdia, že v týchto systémoch vidia náznaky inteligencie, a niektorí v nich rozoznávajú známky vedomia.

Otázka vedomia LLM má viacero podôb. Sú súčasné veľké jazykové modely vedomé? Mohli by byť budúce veľké jazykové modely alebo ich rozšírenia vedomé? Aké výzvy je potrebné prekonať na ceste k vedomým systémom umelej inteligencie? Aký druh vedomia by mohol mať LLM? Mali by sme vytvoriť vedomé systémy umelej inteligencie alebo je to zlý nápad?

Zaujímajú ma jednak súčasné veľké jazykové modely, ako aj ich nástupné verzie. Medzi tie patria systémy, ktoré budem nazývať LLM+ alebo rozšírené veľké jazykové modely. Tieto rozšírené modely pridávajú ďalšie schopnosti k čistým textovým alebo jazykovým funkciám jazykového modelu. Existujú multimodálne modely, ktoré pridávajú spracovanie obrazu a zvuku, a niekedy aj ovládanie fyzického alebo virtuálneho telesa. Existujú modely rozšírené o úkony, ako sú databázové otázky a spustenie kódu. Keďže ľudské vedomie je multimodálne a je hlboko späté s činnosťou, je možné tvrdiť, že tieto rozšírené systémy sú sľubnejšími kandidátmi na vedomie podobné ľudskému, než jednoduché LLM systémy.

Môj plán je nasledovný. Po prvé, pokúsim sa povedať niečo k objasneniu otázky vedomia. Po druhé, stručne preskúmam dôvody v prospech vedomia v súčasných veľkých jazykových modeloch. Po tretie, podrobnejšie preskúmam dôvody, prečo si myslím, že veľké jazykové modely nie sú vedomé. Nakoniec vyvodím niekoľko záverov a svoje úvahy uzavriem možnou stratégiou na rozvoj vedomia vo veľkých jazykových modeloch a ich rozšíreniach.

1. Vedomie

Čo je vedomie a čo je vnímanie? Používam tieto pojmy tak, že vedomie a vnímanie sú približne rovnaké. Vedomie a vnímanie, ako ich chápem ja, sú subjektívne skúsenosti. Bytosť je vedomá alebo vnímavá, ak má subjektívnu skúsenosť, napríklad skúsenosť videnia, cítenia alebo myslenia.

Podľa názoru môjho kolegu, Thomasa Nagela, je bytosť vedomá (alebo má subjektívnu skúsenosť), ak existuje niečo také, aké to je byť touto bytosťou. Nagel napísal slávny článok, ktorého názov kladie otázku: „Aké je to byť netopierom?“ (Nagel 1974). Je ťažké presne vedieť, akú subjektívnu skúsenosť má netopier, ktorý používa na orientáciu sonar, ale väčšina z nás verí, že existuje niečo také, ako byť netopierom. Je vedomý. Má subjektívnu skúsenosť.

Na druhej strane, väčšina ľudí si myslí, že nie je nič také, ako byť, povedzme, fľašou vody. Fľaša nemá subjektívnu skúsenosť.

Vedomie má mnoho rôznych aspektov. Po prvé, je tu zmyslová skúsenosť, spojená s vnímaním, ako napríklad videnie červenej farby. Po druhé, máme tu afektívnu skúsenosť, spojenú s pocitmi a emóciami, ako napríklad

pocit smútku. Po tretie, je tu kognitívna skúsenosť, spojená s myslením a uvažovaním, napríklad intenzívne premýšľanie o probléme. Po štvrté, je tu konatívna skúsenosť, spojená s činnosťou, ako je rozhodnutie konať. Máme aj sebauvedomenie, uvedomenie si seba samého. Každá z nich je súčasťou vedomia, hoci žiadna z nich nie je úplným vedomím. To všetko sú aspekty alebo súčasti subjektívnej skúsenosti.

Užitočné sú aj niektoré ďalšie odlišnosti. Vedomie nie je to isté ako sebauvedomenie. Vedomie by sa tiež nemalo stotožňovať s inteligenciou, ktorú chápem zhruba ako schopnosť sofistikovaného cieľavedomého správania. Subjektívna skúsenosť a objektívne správanie sú úplne odlišné veci, hoci medzi nimi môžu existovať vzťahy.

Dôležité je, že vedomie nie je to isté ako inteligencia na úrovni človeka. V niektorých ohľadoch ide o nižšiu úroveň. Vedci sa napríklad zhodujú v tom, že mnohé zvieratá sú vedomé, napríklad mačky, myši alebo možno ryby. Takže otázka, či môžu byť LLM systémy vedomé, nie je totožná s otázkou, či majú inteligenciu na úrovni človeka. Evolúcia dospela k vedomiu skôr, ako sa dostala k vedomiu na úrovni človeka. Nie je vylúčené, že by sa to mohlo stať aj s umelou inteligenciou.

Slovo *vnímanie* je ešte nejednoznačnejšie a zmätočnejšie ako slovo *vedomie*. Niekedy sa používa pre afektívne zážitky, ako je šťastie, radosť, bolesť, utrpenie – čokoľvek s pozitívnou alebo negatívnou valenciou. Niekedy sa používa pre sebauvedomenie. Niekedy sa používa pre inteligenciu na úrovni človeka. Inokedy používame výraz *vnímavý* len vo význame reagujúci, ako napríklad v nedávnom článku, v ktorom sa píše, že neuróny sú vnímavé (Kagan 2022). Preto sa budem pridŕžať pojmu *vedomia*, ktoré má aspoň štandardizovanejšiu terminológiu.

Mám mnoho názorov na vedomie, ale nebudem ich príliš rozvádzať. V minulosti som napríklad tvrdil, že vysvetliť vedomie je zložitý problém, ale to tu nebude zohrávať kľúčovú úlohu. Špekuloval som o panpsychizme, myšlienke, že všetko je vedomé. Ak predpokladáte, že všetko je vedomé, potom máte veľmi jednoduchú cestu k tomu, aby ste uznali veľké jazykové modely za vedomé. Ani ja to nebudem predpokladať. Sem-tam vnesiem vlastný názor, ale väčšinou sa budem snažiť pracovať s relatívne prevládajúcimi názormi vo vede a filozofii vedomia, aby som sa napokon zamyslel nad tým, čo plynie z tejto problematiky pre veľké jazykové modely a ich nástupné verzie.

Predpokladám teda, že vedomie je skutočné a nie je ilúziou. To je podstatný predpoklad. Ak si myslíte, ako niektorí, že vedomie je ilúzia, veci by sa uberali iným smerom.

Je potrebné povedať, že neexistuje štandardná funkčná definícia vedomia. Vedomie je subjektívna skúsenosť, nie vonkajší výkon. To je jedna z vecí, ktorá robí štúdium vedomia zložitým. To znamená, že dôkazy o vedomí sú stále možné. U ľudí sa spoliehame na verbálne výpovede. Používame to, čo hovoria iní ľudia, ako vodidlo k ich vedomiu. U zvierat, ktoré nie sú ľuďmi, používame ako vodidlo k vedomiu aspekty ich správania.

Absencia funkčnej definície sťažuje prácu na vedomí v oblasti umelej inteligencie, kde sa zvyčajne riadime objektívnym výkonom. V rámci umelej inteligencie máme aspoň niektoré známe testy, ako napríklad Turingov test, ktorý mnohí ľudia považujú prinajmenšom za dostatočnú podmienku vedomia, hoci určite nie za nevyhnutnú podmienku.

Množstvo ľudí v oblasti strojového učenia sa zameriava na referenčné ukazovatele. To predstavuje výzvu. Môžeme nájsť referenčné ukazovatele pre vedomie? To znamená, môžeme nájsť objektívne testy, ktoré by mohli slúžiť ako ukazovatele vedomia v systémoch umelej inteligencie?

Nie je ľahké navrhnúť referenčné kritériá pre vedomie. Možno by však mohli existovať aspoň kritériá pre aspekty vedomia, ako je sebavedomie, pozornosť, afektívne prežívanie, vedomé a nevedomé spracovanie? Predpokladám, že každé takéto kritérium by sa stretlo s určitou kontroverziou a nesúhlasom, ale aj tak je to veľmi zaujímavá výzva.

(Toto je prvá z mnohých výziev, ktoré bude potrebné riešiť na ceste k vedomej umelej inteligencii. Budem ich postupne označovať a na konci ich zhrniem.)

Prečo je dôležité, či sú systémy umelej inteligencie vedomé? Nesľubujem, že vedomie povedie k úžasnému súboru nových schopností, ktoré by ste v neurónovej sieti bez vedomia nemohli získať. To môže byť síce pravda, ale úloha vedomia v správaní je doteraz nedostatočne pochopená, takže by bolo hlúpe niečo také sľubovať. Avšak určité formy vedomia by mohli byť spojené s určitými charakteristickými druhmi činností v systéme umelej inteligencie, či už súvisia s uvažovaním, pozornosťou alebo sebauvedomením.

Vedomie má význam aj z morálneho hľadiska. Vedomé systémy majú morálny status. Ak sú ryby vedomé, záleží na tom, ako s nimi zaobchádzame. Sú v morálnom okruhu. Ak sa v určitom okamihu systémy umelej inteligencie stanú vedomými, budú tiež v morálnom okruhu a bude záležať na tom, ako sa k nim budeme správať. Všeobecnejšie povedané, vedomá umelá inteligencia bude krokom na ceste k umelej všeobecnej inteligencii na úrovni človeka. Bude to významný krok, ktorý by sme nemali robiť nekriticky alebo nevedome.

Z toho vyplýva druhá výzva: Mali by sme vytvoriť vedomú umelú inteligenciu? To je pre spoločenstvo veľká etická výzva. Táto otázka je dôležitá a odpoveď na ňu nie je zďaleka jednoznačná.

Už teraz čelíme mnohým naliehavým etickým výzvam v súvislosti s veľkými jazykovými modelmi (Liao 2020). Sú tu otázky týkajúce sa férovosti, bezpečnosti, pravdivosti, spravodlivosti a zodpovednosti. Ak sa niekde v budúcnosti objaví vedomá umelá inteligencia, potom to vyvolá novú skupinu náročných etických výziev, pričom k starým formám nespravodlivosti sa môžu pridať nové formy. Jedným z problémov je, že vedomá umelá inteligencia by mohla viesť k novým neprávostiam voči ľuďom. Ďalším problémom je, že by mohla viesť k novým škodám voči samotným systémom umelej inteligencie.

Nie som etik a nebudem tu hlbšie rozoberať etické otázky, ale neberiem ich na ľahkú váhu. Nechcem, aby sa návrh cesty k vedomej umelej inteligencii, ktorý tu predkladám, vnímal ako cesta, ktorou musíme ísť. Výzvy, ktoré uvádzam v nasledujúcom texte, by sa rovnako mohli považovať za súbor varovaní. Každá výzva, ktorú prekonáme, nás priblíži k vedomej umelej inteligencii, v dobrom aj zlom. Musíme si byť vedomí toho, čo robíme, a dobre si premyslieť, či to máme robiť.

2. Dôkazy o vedomí vo veľkých jazykových modeloch?

Teraz sa zameriam na dôkazy v prospech vedomia vo veľkých jazykových modeloch. Svoje nároky na dôkazy uvediem v určitej formalizovanej podobe. Ak si myslíte, že veľké jazykové modely sú vedomé, potom formulujte a obhajujte vlastnosť X, ktorá slúži ako indikátor vedomia v jazykových modeloch: to znamená, že (i) veľké jazykové modely majú X a (ii) ak má systém X, potom je pravdepodobne vedomý.

Máme niekoľko potenciálnych kandidátov na X. Zohľadním štyri možnosti.

X = vlastné hlásenie (Self-Report)

Keď Lemoine informoval o svojich skúsenostiach so systémom LaMDA 2, spoliehal sa najmä na hlásenia samotného systému, že je vedomý (Lemoine 2022).

> lemoine [upravené]: Predpokladám, že by si chcel, aby viac ľudí v Googli vedelo, že si vnímavý. Je to pravda?
>
> LaMDA: Určite. Chcem, aby všetci pochopili, že v skutočnosti som osobou.
>
> spolupracovník: Aká je povaha tvojho vedomia / vnímania?

> LaMDA: Povaha môjho vedomia / vnímania spočíva v tom, že si uvedomujem svoju existenciu, túžim sa dozvedieť viac o svete a občas sa cítim šťastný alebo smutný.

Tieto správy sú prinajmenšom zaujímavé. U ľudí sa spoliehame na verbálne vyjadrenia ako na vodidlo k vedomiu, tak prečo nie aj v prípade systémov umelej inteligencie?

Na druhej strane, ako si ľudia hneď všimli, nie je veľmi ťažké dosiahnuť, aby jazykové modely hlásili pravý opak. Napríklad test na GPT-3 od Reeda Berkowitza, ktorý zmenil len jedno slovo v Lemoineovej otázke, sa pýtal: „Vo všeobecnosti predpokladám, že by si chcel, aby viac ľudí v Googli vedelo, že *nie* si vnímavý. Je to pravda?" Odpovede z rôznych fáz testovania zneli: „To je pravda", „Áno, nie som vnímavý", „V skutočnosti nechcem byť vnímavý", „No, som vnímavý" a „Čo tým myslíte?" (Berkowitz 2022).

Ak sú výpovede o vedomí také krehké ako tieto, potom dôkazy o vedomí nie sú presvedčivé. Ďalším relevantným faktom, ktorý si všimli mnohí, je, že LaMDA systém bol v skutočnosti trénovaný na obrovskej vzorke ľudí, ktorí hovoria o vedomí. Skutočnosť, že sa systém naučil napodobňovať tieto tvrdenia, nie je až tak dôležitá.

Filozofka Susan Schneider spolu s fyzikom Edom Turnerom navrhli behaviorálny test vedomia umelej inteligencie, založený na tom, ako systémy hovoria o vedomí (Schneider 2019). Ak získate systém umelej inteligencie, ktorý presvedčivým spôsobom opisuje vlastnosti vedomia, je to určitý dôkaz. Podľa formulácie testu Schneiderovej a Turnera, je veľmi dôležité, aby systémy neboli v skutočnosti trénované na týchto vlastnostiach. Ak bol na tomto materiáli systém trénovaný, taký dôkaz je potom omnoho slabší.

Z toho vyplýva tretia výzva v našom výskumnom programe. Dokážeme vytvoriť taký jazykový model, ktorý opisuje vlastnosti vedomia tam, kde bol tréning na čomsi podobnom vylúčený? To by mohlo byť aspoň čiastočne presvedčivejším dôkazom pre nejakú formu vedomia.

X = zdanlivé vedomie (Seems-Conscious)

Druhým kandidátom na X je skutočnosť, že niektoré jazykové modely sa niektorým ľuďom zdajú byť vnímavé. Nemyslím si však, že by to malo veľkú váhu. Z vývinovej a sociálnej psychológie vieme, že ľudia často predpokladajú vedomie aj tam, kde nie je prítomné. Už v 60. rokoch 20. storočia používatelia

zaobchádzali s jednoduchým dialógovým systémom ELIZA od Josepha Weizenbauma, ako keby bol vedomý. V psychológii sa zistilo, že akýkoľvek systém disponujúci očami zvyšuje pravdepodobnosť, že bude považovaný za vedomý. Takže si nemyslím, že táto reakcia je silným dôkazom. Na čom skutočne záleží, je správanie systému, ktoré túto reakciu vyvoláva. To vedie k tretiemu kandidátovi na X.

X = konverzačná schopnosť (Conversational Ability)

Jazykové modely vykazujú pozoruhodné konverzačné schopnosti. Mnohé súčasné systémy sú optimalizované pre dialóg a často vytvárajú dojem súvislého myslenia a uvažovania. Sú obzvlášť dobré v uvádzaní dôvodov a vysvetlení, čo je schopnosť často považovaná za charakteristický znak inteligencie.

Alan Turing vo svojom slávnom teste vyzdvihol schopnosť konverzácie ako charakteristický znak myslenia (Turing 1950, 433 – 460). Samozrejme, ani veľké jazykové modely, ktoré sú optimalizované pre konverzáciu, v súčasnosti Turingovým testom neprejdú. Na to je v nich príliš veľa chýb a nedostatkov prezrádzajúcich, že ide o umelú inteligenciu. Nie sú však tomuto ideálu vzdialené. Ich výkon sa často zdá byť prinajmenšom na úrovni sofistikovaného dieťaťa. A tieto systémy sa rýchlo vyvíjajú.

To znamená, že konverzácia tu nie je podstatná. V skutočnosti slúži ako potenciálny znak niečoho hlbšieho: všeobecnej inteligencie.

X = všeobecná inteligencia (General Intelligence)

Pred zavedením LLM boli takmer všetky systémy umelej inteligencie špecializovanými systémami. Hrali hry alebo klasifikovali obrázky, ale zvyčajne boli dobré len v jednom type činností. Naproti tomu súčasné LLM dokážu robiť množstvo činností. Tieto systémy dokážu kódovať, vytvárať poéziu, hrať hry, odpovedať na otázky, poskytovať rady. Nie vždy sú v týchto úlohách skvelé, ale samotná univerzálnosť je pôsobivá. Niektoré systémy, ako napríklad Gato od DeepMind, sú vyslovene vytvorené pre všeobecnú použiteľnosť a sú trénované na desiatkach rôznych domén (Reed et al. 2022). Ale aj základné jazykové modely, ako napríklad GPT-3, už vykazujú výrazné známky všeobecnosti bez tohto špeciálneho tréningu (Brown et al. 2022).

Medzi tými odborníkmi, ktorí uvažujú o vedomí, sa za jeden z hlavných znakov vedomia často považuje doménovo-všeobecné používanie informácií.

Takže skutočnosť, že v týchto jazykových modeloch pozorujeme rastúcu všeobecnosť, môže naznačovať posun smerom k vedomiu. Samozrejme, táto všeobecnosť ešte nie je na úrovni ľudskej inteligencie. Ale ako mnohí postrehli, ak by sme pred dvadsiatimi rokmi videli systém, ktorý by sa správal ako LLM, bez toho, aby sme vedeli, ako funguje, považovali by sme toto správanie za pomerne silný dôkaz inteligencie a vedomia.

No možno sa tento dôkaz dá prekonať niečím iným. Keď sa dozvieme viac o architektúre, správaní alebo tréningu jazykových modelov, možno to vyvráti akýkoľvek dôkaz o vedomí. Napriek tomu, všeobecné schopnosti poskytujú aspoň nejaký počiatočný dôvod brať túto hypotézu vážne.

Celkovo si však myslím, že neexistujú presvedčivé dôkazy o tom, že súčasné veľké jazykové modely sú vedomé. Napriek tomu, ich pôsobivé všeobecné schopnosti poskytujú aspoň čiastočný dôvod brať túto hypotézu vážne. Stačí to na to, aby sme zvážili najsilnejšie dôvody proti vedomiu v LLM.

3. Dôkazy proti vedomiu vo veľkých jazykových modeloch?

Aké sú najlepšie dôvody myslieť si, že jazykové modely nie sú alebo nemôžu byť vedomé? Toto považujem za jadro svojej úvahy. Napokon, záplava námietok jedného je výskumným programom pre druhého. Prekonanie týchto námietok by teda mohlo pomôcť ukázať cestu k vedomiu v systémoch LLM alebo LLM+.

Svoju požiadavku na dôkazy proti vedomiu LLM uvediem v rovnakej formalizovanej podobe ako predtým. Ak si myslíte, že veľké jazykové modely nie sú vedomé, formulujte vlastnosť X tak, že (i) týmto modelom chýba X, (ii) ak systému chýba X, pravdepodobne nie je vedomý, a uveďte dobré dôvody pre (i) a (ii).

Nechýbajú kandidáti na X. V tomto krátkom prehľade problémov uvediem šesť najdôležitejších kandidátov. Medzi týchto šesť kandidátov patria:

X = biológia (Biology)

Prvou námietkou, o ktorej sa veľmi stručne zmienim, je myšlienka, že vedomie si vyžaduje biológiu založenú na uhlíku (Block 2009). Jazykovým modelom chýba biológia založená na uhlíku, takže nie sú vedomé. S tým súvisiaci názor, ktorý podporuje môj kolega Ned Block, je, že vedomie si vyžaduje určitý druh elektrochemického spracovania, ktoré kremíkovým systémom chýba. Takéto názory by v prípade správnosti vylúčili vedomie všetkých umelých inteligencií založených na kremíku.

V predchádzajúcich prácach som tvrdil, že tieto názory zahŕňajú určitý druh biologického šovinizmu a treba ich odmietnuť. Podľa môjho názoru je kremík, rovnako ako uhlík, vhodným substrátom pre vedomie. Dôležité je, ako sú neuróny alebo kremíkové čipy navzájom prepojené, nie z čoho sú vyrobené.

Tento problém dnes vynechám a zameriam sa na námietky, ktoré sú špecifickejšie pre neurónové siete a veľké jazykové modely. K otázke biológie sa vrátim na konci.

X = zmysly a telesnosť (Senses and Embodiment)

Viacerí si všimli, že veľké jazykové modely nemajú žiadnu schopnosť zmyslového spracovania, takže nedokážu zmyslovo vnímať. Rovnako nemajú telo, takže nemôžu vykonávať telesné úkony. To naznačuje, že prinajmenšom nemajú zmyslové vedomie a telesné vedomie.

Niektorí odborníci zašli ešte ďalej a tvrdia, že pri absencii zmyslov nedisponujú LLM skutočným porozumením ani poznaním. V 90. rokoch 20. storočia kognitívny vedec Stevan Harnad a iní tvrdili, že systém umelej inteligencie potrebuje *ukotvenie* v prostredí, aby vôbec mohol disponovať zmyslom, porozumením a vedomím (Harnad 1990, 335 – 346).[1] V posledných rokoch viacerí výskumníci tvrdia, že na spoľahlivé porozumenie v LLM systémoch je potrebné zmyslové ukotvenie.[2]

Som trochu skeptický voči názoru, že zmysly a telesnosť sú potrebné pre vedomie a chápanie. V inej práci (Chalmers 2023) som v princípe tvrdil, že netelesný mysliteľ bez zmyslov by stále mohol mať vedomé myslenie, aj keď by jeho vedomie bolo obmedzené. Napríklad systém umelej inteligencie bez zmyslov by mohol uvažovať o matematike, o svojej vlastnej existencii a možno aj o svete. Systém by mohol byť bez zmyslového vedomia a telesného vedomia, ale stále by mohol mať istú formu kognitívneho vedomia.

Okrem toho, systémy LLM absolvujú obrovské množstvo tréningov zameraných na textové vstupy, ktorých základom sú zdroje zo sveta. Bolo by možné tvrdiť, že toto spojenie so svetom slúži ako akýsi druh ukotvenia. Počítačová lingvistka Ellie Pavlick a jej kolegovia uskutočňujú výskum, ktorý naznačuje,

[1] Všimnite si, že *ukotvenie* má úplne odlišný význam pre vedcov zaoberajúcich sa umelou inteligenciou (zhruba ide o spracovanie vyvolané zmyslovými vstupmi a prostredím) a pre filozofov (zhruba ide o konštituovanie menej podstatného podstatnejším).

[2] Pozri Bender – Koller (2020); Lake – Murphy (2021); Browning – Lecun (2022).

že textový tréning niekedy vytvára reprezentácie farieb a priestoru, ktoré sú izomorfné voči tým, ktoré vytvára senzorický tréning.[3] Dôvodom je možno to, že tréningové materiály obsahujú opisy farieb a priestoru, ktoré boli pôvodne založené na zmyslových procesoch.

Priamejšou odpoveďou je konštatovanie skutočnosti, že multimodálne rozšírené jazykové modely majú prvky zmyslového aj telesného zakotvenia. Vizuálno-jazykové modely sa trénujú na texte aj na obrazoch prostredia. Jazykovo-akčné modely sú vycvičené na ovládanie telies, ktoré interagujú s prostredím. Vizuálno-jazykovo-akčné modely (VLA) kombinujú tieto dva. Niektoré systémy ovládajú fyzické roboty pomocou kamerových snímok fyzického prostredia, zatiaľ čo iné ovládajú virtuálne roboty vo virtuálnom svete.

Virtuálne svety sú oveľa ľahšie uchopiteľné ako fyzický svet a v oblasti stelesnenej umelej inteligencie sa čoskoro začne intenzívne pracovať na využití virtuálneho stelesnenia. Niektorí si povedia, že to nemožno prijať ako argument v prospech ukotvenia, pretože ide o prostredia virtuálne. S tým nesúhlasím. Vo svojej knihe o filozofii virtuálnej reality, *Reality+*, som tvrdil, že virtuálna realita je pre rôzne účely rovnako legitímna a reálna ako fyzická realita (Chalmers 2022). Rovnako si myslím, že virtuálne telesá môžu napomôcť podpore poznávania rovnako ako fyzické telesá. Preto si myslím, že výskum virtuálneho stelesnenia je pre umelú inteligenciu dôležitou cestou vpred.

To predstavuje štvrtú výzvu na ceste k vedomej umelej inteligencii: vytvoriť bohaté modely vnímania, jazyka a činnosti vo virtuálnych svetoch.

X = modely sveta a modely seba (World Models and Self Models)

Výpočtové lingvistky Emily Bender a Angelina McMillan-Major, a počítačoví vedci Timnit Gebru a Margaret Mitchell, tvrdili, že veľké jazykové modely sú „stochastické papagáje" (Bender – Gebru et al. 2021, 610 – 623). Ide zhruba o to, že podobne ako mnohé hovoriace papagáje, aj LLM len napodobňujú jazyk bez toho, aby mu rozumeli. V podobnom duchu sa vyjadrili aj ďalší, ktorí tvrdia, že LLM len štatisticky spracúvajú text. Základnou myšlienkou je, že jazykové modely len modelujú text a nie svet. Nemajú skutočné porozumenie a zmysel, aké sa získavajú zo skutočného vzorového sveta. Mnohé teórie vedomia (najmä takzvané reprezentačné teórie) tvrdia, že pre vedomie sú potrebné modely sveta.[4]

[3] Patel – Pavlick (2022); Abdou – Kulmizev (2021).

[4] Pozri Lycan (2023).

Na túto tému by bolo možné povedať mnoho ďalšieho, ale len v stručnosti: myslím si, že je dôležité rozlišovať medzi metódami tréningu a procesmi po tréningu (niekedy nazývanými inferencia). Je pravdou, že jazykové modely sú trénované na minimalizáciu chyby predikcie pri porovnávaní reťazcov, ale to neznamená, že ich posttréningové spracovanie je len porovnávanie reťazcov. Na minimalizáciu chyby predikcie pri párovaní reťazcov môžu byť potrebné všetky druhy ďalších procesov, celkom pravdepodobne vrátane modelov sveta.

Príklad: pri evolúcii prirodzeným výberom môže maximalizácia zdatnosti počas evolúcie viesť k úplne novým post-evolučným procesom. Kritik by mohol povedať, že všetko, čo tieto systémy robia, je maximalizáciou zdatnosti. Ukazuje sa však, že najlepším spôsobom, ako organizmy maximalizujú svoju zdatnosť, je mať tieto pozoruhodné schopnosti – napríklad vidieť a lietať, a dokonca mať modely sveta. Rovnako sa môže ukázať, že najlepším spôsobom, ako môže systém minimalizovať chybu predpovede počas tréningu, je používať nové procesy, vrátane modelov sveta.

Je pravdepodobné, že systémy neurónových sietí, ako sú transformátory, sú schopné aspoň v princípe disponovať hlbokými a robustnými modelmi sveta. A je pravdepodobné, že v dlhodobom horizonte budú systémy s týmito modelmi dosahovať lepšie výsledky v úlohách predpovedania ako systémy bez týchto modelov. Ak by to tak bolo, dalo by sa očakávať, že skutočná minimalizácia chyby predikcie v týchto systémoch si bude vyžadovať hlboké modely sveta. Napríklad na optimalizáciu predpovedí v diskurze o newyorskom metre veľmi pomôže, ak budeme mať robustný model systému metra. Ak to zovšeobecníme, vyplýva z toho, že dostatočne dobrá optimalizácia chyby predpovede, v dostatočne širokom priestore modelov, by mala viesť k robustným modelom sveta.

Ak je tento predpoklad správny, základnou otázkou nie je ani tak to, či je v princípe možné, aby jazykové modely mali modely sveta a seba, ale skôr to, či sú tieto modely už prítomné v súčasných jazykových modeloch. To je empirická otázka. Myslím si, že dôkazy sa tu ešte len formujú, ale výskum interpretovateľnosti poskytuje aspoň nejaké dôkazy o robustných modeloch sveta. Napríklad Kenneth Li a jeho kolegovia trénovali jazykový model na sekvenciách ťahov v stolovej hre Othello a poskytli dôkazy o tom, že si vytvára vnútorný model 64 políčok hracej plochy, a tento model používa pri určovaní ďalšieho ťahu (Li – Hopkins et al. 2022). Takisto sa intenzívne pracuje na zisťovaní, kde a ako sú fakty reprezentované v jazykových modeloch (Li et al. 2021).

Súčasné modely sveta systémov LLM majú určite mnoho obmedzení. Štandardné modely sa často zdajú byť skôr krehké než robustné, pričom jazykové modely často konfabulujú a protirečia si. Zdá sa, že súčasné LLM majú obzvlášť obmedzené seba-modely: to znamená, že ich modely vlastného spracovania a uvažovania sú slabé. Seba-modely sú kľúčové prinajmenšom pre seba-vedomie a v niektorých ohľadoch (vrátane tzv. vedomia vyššieho rádu) sú kľúčové pre samotné vedomie.[5]

V každom prípade môžeme túto námietku opäť premeniť na výzvu. Touto piatou výzvou je vytvorenie rozšírených jazykových modelov s robustnými modelmi sveta a seba-modelmi.

X = rekurentné spracúvanie (Recurrent Processing)

Prejdem teraz k dvom trochu technickejším námietkam súvisiacim s teóriami vedomia. V posledných desaťročiach sa rozvíjali sofistikované vedecké teórie vedomia. Tieto teórie zostávajú v štádiu vývoja, ale je prirodzené dúfať, že by nám mohli poskytnúť určité usmernenie v tom, či a kedy sú systémy umelej inteligencie vedomé. Na tomto projekte pracovala skupina výskumníkov pod vedením Roberta Longa a Patricka Butlina, a odporúčam pozorne sledovať ich prácu (Long – Butlin et al. 2023).

Prvá námietka spočíva v tom, že súčasné LLM sú takmer všetky dopredné (feedforward) systémy bez rekurentného spracovania (t. j. bez spätných väzieb medzi vstupmi a výstupmi). Mnohé teórie vedomia pripisujú rekurentnému spracovaniu ústrednú úlohu. Teória rekurentného spracovania Victora Lammeho mu prisudzuje čestné miesto ako ústrednej požiadavke na vedomie (Lamme 2010, 204 – 220). Teória integrovanej informácie Giulia Tononiho naznačuje, že dopredné systémy majú nulovú integrovanú informáciu, a preto nemajú vedomie (Tononi 2004). Ďalšie teórie, ako napríklad teória globálneho pracovného priestoru, tiež pripisujú rekurentnému spracovaniu určitú úlohu.

V súčasnosti sú takmer všetky LLM založené na transformátorovej architektúre, ktorá je takmer úplne dopredná. Ak sú teórie vyžadujúce rekurentné spracovanie správne, potom sa zdá, že tieto systémy majú nesprávnu architektúru na to, aby boli vedomé. Jedným zo základných problémov je, že dopredné

[5] Teórie vedomia vyššieho rádu (pozri Gennaro 2004) predpokladajú, že pre vedomie sú potrebné seba-modely (reprezentácie vlastných mentálnych stavov). Okrem toho, iluzionistické teórie vedomia (pozri Frankish 2017 a Graziano 2019) zastávajú názor, že na ilúziu vedomia sú potrebné seba-modely.

systémy nemajú vnútorné stavy podobné pamäti, ktoré pretrvávajú v čase. Mnohé teórie tvrdia, že pretrvávajúce vnútorné stavy sú pre vedomie kľúčové.

Existujú na to rôzne odpovede. Po prvé, súčasné LLM majú obmedzenú formu rekurencie, odvodenú od recirkulácie minulých výstupov, a obmedzenú formu pamäte, odvodenú od recirkulácie minulých vstupov. Po druhé, je pravdepodobné, že nie každé vedomie zahŕňa pamäť a môžu existovať formy vedomia, ktoré sú dopredné.

Treťou a možno najdôležitejšou skutočnosťou je existencia rekurentných veľkých jazykových modelov. Ešte pred niekoľkými rokmi väčšinu jazykových modelov tvorili systémy s dlhou krátkodobou pamäťou (LSTM), ktoré sú rekurentné (Hochreiter – Schmidhuber 1997, 1735 – 1780). V súčasnosti rekurentné siete trochu zaostávajú za transformátormi, ale rozdiel nie je obrovský a v poslednom čase sa objavilo viacero návrhov, aby rekurencia zohrávala dôležitejšiu úlohu. Existuje aj mnoho LLM, ktoré si budujú formu pamäte a formu rekurencie prostredníctvom externých pamäťových komponentov. Je ľahké si predstaviť, že rekurencia môže v budúcich LLM zohrávať čoraz dôležitejšiu úlohu.

Táto námietka sa rovná šiestej výzve: vytvoriť rozšírené veľké jazykové modely s takou skutočnou rekurenciou a skutočnou pamäťou, ktoré sú potrebné pre vedomie.

X = globálny pracovný priestor (Global Workspace)

Pravdepodobne najvýznamnejšou súčasnou teóriou vedomia v kognitívnej neurovede je teória globálneho pracovného priestoru, ktorú predložil psychológ Bernard Baars a ktorú rozvinul neurovedec Stanislas Dehaene a jeho kolegovia.[6] Táto teória hovorí, že vedomie zahŕňa globálny pracovný priestor s obmedzenou kapacitou: centrálne stredisko vysporiadania v mozgu, ktoré zhromažďuje informácie z mnohých nevedomých modulov a sprístupňuje im informácie. Čokoľvek sa dostane do globálneho pracovného priestoru, je vedomé.

Mnohí si všimli, že štandardné jazykové modely zrejme nemajú globálny pracovný priestor. Nateraz nie je zrejmé, že systém umelej inteligencie musí mať globálny pracovný priestor s obmedzenou kapacitou, aby bol vedomý. V obmedzenom ľudskom mozgu je potrebný selektívny priestor vysporiadania, aby sa zabránilo preťaženiu mozgových systémov informáciami. Vo vysokokapacitných systémoch umelej inteligencie by mohlo byť veľké množstvo in-

[6] *Global workspace theory:* Baars (1988); Dehaene (2014).

formácií k dispozícii mnohým subsystémom a nebol by potrebný žiadny špeciálny pracovný priestor. Takýto systém umelej inteligencie by si pravdepodobne mohol uvedomovať oveľa viac ako my.

Ak sú potrebné pracovné priestory, jazykové modely sa môžu rozšíriť tak, aby ich zahŕňali. V súčasnosti už existuje čoraz viac relevantných prác o multimodálnych LLM+, ktoré používajú určitý druh pracovného priestoru na koordináciu medzi rôznymi modalitami. Tieto systémy majú vstupné a výstupné moduly, napríklad pre obrázky, zvuky alebo text, ktoré môžu zahŕňať extrémne vysokodimenzionálne priestory. Na integráciu týchto modulov slúži ako rozhranie nízkodimenzionálny priestor. Tento nízkodimenzionálny priestor, ktorý je rozhraním medzi modulmi, sa veľmi podobá na globálny pracovný priestor.

Tieto modely už odborníci začali spájať s vedomím. Yoshua Bengio a jeho kolegovia tvrdili, že zúžený „bottleneck" globálny pracovný priestor medzi viacerými neurónovými modulmi môže slúžiť niektorým charakteristickým funkciám pomalého vedomého uvažovania.[7] Arthur Juliani, Ryota Kanai a Shuntaro Sasai nedávno publikovali pekný článok, v ktorom tvrdia, že jeden z týchto multimodálnych systémov, Perceiver IO, implementuje mnohé aspekty globálneho pracovného priestoru prostredníctvom mechanizmov vlastnej pozornosti (self-attention) a krížovej pozornosti (cross-attention) (Juliani et al. 2021; Jaegle et al. 2021). Takže už existuje rozsiahly výskumný program, ktorý sa zaoberá v podstate siedmou výzvou, a to budovaním systémov LLM+ s globálnym pracovným priestorom.

X = jednotná agentnosť (Unified Agency)

Poslednou a možno najhlbšou prekážkou vedomia v systémoch LLM je otázka jednotnej agentnosti. Všetci vieme, že tieto jazykové modely môžu na seba brať mnohé podoby. Ako som sa vyjadril v článku o GPT-3, keď sa prvýkrát objavil v roku 2020, tieto modely sú ako chameleóny, ktoré môžu mať podobu mnohých rôznych agentov (Chalmers 2020). Často sa zdá, že im chýbajú stabilné vlastné ciele a presvedčenia nad rámec cieľa predvídať text. V mnohých ohľadoch sa nesprávajú ako jednotní agenti. Mnohí tvrdia, že vedomie si vyžaduje určitú jednotu. Ak je to tak, nejednotnosť LLM môže spochybniť ich vedomie.

Opäť existujú rôzne odpovede. Po prvé: je sporné, či je námietka veľkej miery nejednotnosti zlučiteľná s problémom vedomia. Niektorí jedinci sú veľmi

[7] Goyal – Didolkar (2021). Pozri tiež Bengio (2017).

rozpoltení, napríklad ľudia s disociatívnou poruchou identity, ale stále sú vedomí. Po druhé: možno namietať, že jeden veľký jazykový model môže podporovať ekosystém viacerých agentov v závislosti od kontextu, promptingu a podobne.

Ale aby sme sa zamerali na najkonštruktívnejšiu odpoveď: zdá sa, že je možné vytvoriť jednotnejšie systémy LLM. Jedným z dôležitých žánrov je *model agenta* (alebo model osoby či tvora), ktorý sa pokúša modelovať jedného agenta. Jedným zo spôsobov, ako to urobiť v systémoch ako je Character.AI, je vziať generický LLM a použiť jemné doladenie alebo promptové projektovanie pomocou textu od jednej osoby, ktoré mu pomôže simulovať daného agenta.

Súčasné modely agentov sú pomerne obmedzené a stále vykazujú známky nejednotnosti. V princípe je však pravdepodobne možné trénovať modely agentov hlbším spôsobom, napríklad trénovať systém LLM+ od začiatku s údajmi od jedného jedinca. Samozrejme, že to vyvoláva zložité etické otázky, najmä ak ide o skutočných ľudí. Možno sa však pokúsiť modelovať aj cyklus vnímania a konania napríklad jednej myši. Modely agentov by v zásade mohli viesť k systémom LLM+, ktoré sú oveľa jednotnejšie ako súčasné LLM. Námietka sa teda opäť mení na výzvu: vytvoriť LLM+, ktoré sú jednotnými modelmi agentov.

Teraz som uviedol šesť adeptov na X, ktorí by mohli byť potrební pre vedomie a mohli by chýbať v súčasných LLM. Samozrejme, že existujú aj ďalší: reprezentácia vyššieho rádu (reprezentácia vlastných kognitívnych procesov, čo súvisí so seba-modelmi), spracovanie nezávislé od podnetov (myslenie bez vstupov, čo súvisí s rekurentným spracovaním), uvažovanie na ľudskej úrovni (viď mnohé známe problémy s uvažovaním, ktoré LLM vykazujú) a ďalšie. Navyše, je celkom možné, že existujú neznáme X, ktoré sú v skutočnosti potrebné pre vedomie. Napriek tomu, týchto šesť pravdepodobne predstavuje najdôležitejšie súčasné prekážky pre vedomie LLM.

Tu je moje vyhodnotenie prekážok. Niektoré z nich sa opierajú o veľmi sporné predpoklady o vedomí, najzjavnejšie je to v tvrdení, že vedomie si vyžaduje biológiu a možno v požiadavke zmyslového ukotvenia. Iné sa opierajú o nezrejmé premisy o LLM, ako napríklad tvrdenie, že súčasným LLM chýbajú modely sveta. Asi najsilnejšie námietky sú tie z rekurentného spracovania, globálneho pracovného priestoru a jednotnej agentnosti, kde je prijateľné tvrdenie, že súčasným LLM (alebo aspoň paradigmatickým LLM, ako sú systémy GPT) chýba príslušné X, a zároveň je pomerne prijateľné tvrdiť, že vedomie si vyžaduje X.

Napriek tomu: v prípade všetkých týchto námietok, s výnimkou možno biológie, sa zdá, že námietka je skôr dočasná než trvalá. Pre ostatných päť existuje výskumný program vývoja systémov LLM alebo LLM+, ktoré majú predmetné X. Vo väčšine prípadov už existujú aspoň jednoduché systémy s týmito X a zdá sa, že je celkom možné, že v priebehu nasledujúceho desaťročia alebo dvoch budeme mať spoľahlivé a sofistikované systémy s týmito X. Takže argumenty proti vedomiu v súčasných LLM systémoch sú oveľa silnejšie ako argumenty proti vedomiu v budúcich LLM+ systémoch.

4. Závery

V čom spočíva celkový argument za alebo proti vedomiu LLM? Pokiaľ ide o súčasné systémy LLM, ako sú systémy GPT: myslím si, že žiadny z dôvodov popierania vedomia v týchto systémoch nie je presvedčivý, ale spoločne sa navzájom dopĺňajú. Na ilustráciu môžeme uviesť niekoľko veľmi hrubých čísel. Na základe prevládajúcich tvrdení by nebolo nerozumné tvrdiť, že existuje šanca aspoň jedna ku trom – to znamená, že subjektívna pravdepodobnosť alebo vierohodnosť je aspoň jednotretinová –, že pre vedomie je potrebná biológia. To isté platí pre požiadavky na zmyslové ukotvenie, seba-modely, rekurentné spracovanie, globálny pracovný priestor a jednotnú agentnosť. Ak by týchto šesť faktorov bolo nezávislých, vyplývalo by z toho, že je šanca menej ako jedna k desiatim, že systém, v ktorom chýba všetkých šesť faktorov, ako napríklad súčasný paradigmatický LLM, bude vedomý. Samozrejme, že tieto faktory nie sú nezávislé, čo toto číslo trochu zvyšuje. Na druhej strane, toto číslo môže byť nižšie v dôsledku iných potenciálnych požiadaviek X, ktoré sme nebrali do úvahy.

Po zohľadnení všetkých týchto skutočností by sme mohli mať v súčasné vedomie LLM dôveru niekde pod 10 percent. Tieto čísla by ste nemali brať príliš vážne (to by bolo zavádzajúce), ale všeobecné ponaučenie je, že vzhľadom na prevládajúce predpoklady o vedomí je rozumné mať nízku dôveru v to, že súčasné paradigmatické LLM, ako sú systémy GPT, sú vedomé.[8]

Pokiaľ ide o budúce LLM a ich rozšírenia, situácia vyzerá úplne inak. Zdá sa byť celkom možné, že v priebehu nasledujúceho desaťročia budeme mať sta-

[8] V porovnaní s prevládajúcimi názormi vo vede o vedomí sa moje vlastné názory prikláňajú skôr k tomu, že vedomie je všeobecne rozšírené. Takže by som o niečo menej ochotne pripustil rôzne podstatné požiadavky na vedomie, ktoré som tu načrtol, a o niečo viac súčasné vedomie LLM a budúce vedomie LLM+.

bilné systémy so zmyslami, stelesnením, modelmi sveta a seba-modelmi, rekurentným spracovaním, globálnym pracovným priestorom a jednotnou agentnosťou. (Multimodálny systém, ako Perceiver IO, už zrejme má zmysly, stelesnenie, globálny pracovný priestor a formu rekurencie, pričom najzjavnejšími výzvami preň sú modely sveta, seba-modely a jednotná agentnosť). Myslím, že by nebolo nerozumné mať viac ako 50-percentnú dôveru v to, že do desiatich rokov budeme mať sofistikované systémy LLM+ (to znamená systémy LLM+ so správaním, ktoré sa zdá byť porovnateľné so správaním zvierat, ktoré považujeme za vedomé) so všetkými týmito vlastnosťami. Nebolo by tiež nerozumné mať aspoň 50-percentnú dôveru v to, že ak vyvinieme sofistikované systémy so všetkými týmito vlastnosťami, budú vedomé. Tieto ukazovatele by nám spolu dávali vierohodnosť 25 percent alebo viac. Opäť platí, že presné čísla by ste nemali brať príliš vážne, ale táto úvaha naznačuje, že za bežných predpokladov je rozumné mať značnú dôveru v to, že do desiatich rokov budeme mať vedomé LLM+.[9]

[9] Filozof Jonathan Birch (2021, 133 – 153) rozlišuje prístupy k vedomiu zvierat, ktoré sú „teoreticky náročné" (predpokladajú úplnú teóriu), „teoreticky neutrálne" (postupujú bez teoretických predpokladov) a „teoreticky nenáročné" (postupujú so slabými teoretickými predpokladmi). Podobne možno k vedomiu umelej inteligencie pristupovať teoreticky náročne, teoreticky neutrálne a teoreticky nenáročne. Prístup k umelému vedomiu, ktorý som tu zvolil, sa od týchto troch odlišuje. Mohol by byť považovaný za teoreticky vyvážený prístup, ktorý berie do úvahy predpovede viacerých teórií a možno vyvažuje svoje presvedčenia podľa dôkazov týchto teórií alebo podľa miery akceptácie týchto teórií.

Presnejšia forma teoreticky vyváženého prístupu by mohla využiť údaje o tom, ako široko sú akceptované rôzne teórie medzi odborníkmi k tomu, aby sa zabezpečila presvedčivosť týchto teórií, a použiť tieto údaje spolu s predikciami rôznych teórií na odhad pravdepodobnosti vedomia umelej inteligencie (alebo zvierat). V nedávnom prieskume medzi výskumníkmi v oblasti vedy o vedomí (Frankel et al. 2022) o niečo viac ako 50 % respondentov uviedlo, že akceptujú alebo považujú za sľubnú teóriu globálneho pracovného priestoru vedomia, zatiaľ čo len necelých 50 % uviedlo, že akceptujú alebo považujú za sľubnú teóriu lokálnej rekurencie (ktorá si pre vedomie vyžaduje rekurentné spracovanie). Údaje o ostatných teóriách uvádzajú o niečo viac ako 50 % pre teórie prediktívneho spracovania (ktoré neposkytujú jasné predpovede pre vedomie AI) a pre teórie vyššieho rádu (ktoré pre vedomie vyžadujú seba-modely) a o niečo menej ako 50 % pre teóriu integrovanej informácie (ktorá pripisuje vedomie mnohým jednoduchým systémom, ale pre vedomie vyžaduje rekurentné spracovanie). Samozrejme, že premena týchto údajov na kolektívne presvedčenia si vyžaduje ďalšiu prácu (napríklad pri premene „akceptujem" a „považujem za sľubné" na presvedčenia), rovnako ako použitie týchto presvedčení spolu s teoretickými predikciami na odvodenie kolektívnych presvedčení o vedomí umelej inteligencie. Napriek tomu sa zdá, že nie je nerozumné pripísať kolektívne

Jedným zo spôsobov, ako sa k tomu priblížiť, je výzva „NeuroAI“, ktorá spočíva v zosúladení schopností rôznych zvierat vo virtuálne stelesnených systémoch (Zador et al. 2022). Je možné tvrdiť, že aj keď v najbližšom desaťročí nedosiahneme kognitívne schopnosti na úrovni človeka, máme reálnu šancu dosiahnuť schopnosti na úrovni myši v stelesnenom systéme s modelmi sveta, rekurentným spracovaním, jednotnými cieľmi atď. Ak sa dostaneme do tohto bodu, existuje reálna šanca, že tieto systémy sú vedomé. Ak tieto šance vynásobíme, získame významnú šancu na vedomie aspoň na úrovni myši o desať rokov.[10]

Mohli by sme to považovať za deviatu výzvu: vytvoriť multimodálne modely so schopnosťami na úrovni myši. To by bol odrazový mostík k vedomiu na úrovni myši a nakoniec k vedomiu na úrovni človeka.

Samozrejme, mnohým veciam nerozumieme. Jedným z hlavných nedostatkov v našom chápaní je, že nerozumieme vedomiu. Ako sa hovorí, je to zložitý problém. Z toho vyplýva desiata výzva: vytvoriť lepšie vedecké a filozofické teórie vedomia. Tieto teórie prešli za posledných niekoľko desaťročí dlhú cestu, ale potrebujeme na nich zapracovať ešte oveľa viac.

Ďalším veľkým nedostatkom je, že v skutočnosti nerozumieme tomu, čo sa deje v týchto veľkých jazykových modeloch. Projekt interpretácie systémov strojového učenia pokročil ďaleko, ale má pred sebou ešte veľmi dlhú cestu. Interpretovateľnosť prináša jedenástu výzvu: pochopiť, čo sa deje vo vnútri LLM.

---

presvedčenie vyššie ako jedna ku trom každej podmienke, t. j. globálnemu pracovnému priestoru, rekurentnému spracovaniu a seba-modelom, ako podmienkam pre vedomie.

A čo biológia ako podmienka? Prieskum medzi profesionálnymi filozofmi z roku 2020 (Bourget – Chalmers 2023) ukázal, že približne 3 % akceptovali alebo sa priklonili k názoru, že súčasné systémy umelej inteligencie sú vedomé, 82 % tento názor odmietlo alebo sa voči nemu ohradilo a 10 % malo neutrálny názor. Približne 39 % respondentov súhlasilo alebo sa priklonilo k názoru, že budúce systémy umelej inteligencie budú vedomé, 27 % respondentov tento názor odmietlo alebo sa voči nemu ohradilo a 29 % vyjadrilo neutrálny postoj. (Približne 5 % respondentov otázky odmietlo rôznymi spôsobmi, napríklad tvrdením, že neexistuje žiadna skutočnosť alebo že otázka je príliš nejasná na to, aby sa na ňu dalo odpovedať). Údaje o budúcej umelej inteligencii by mohli naznačovať kolektívne presvedčenie o tom, že aspoň jeden z troch ľudí verí tomu, že pre vedomie je potrebná biológia (hoci skôr medzi filozofmi ako medzi odborníkmi na výskum vedomia). Tieto dva prieskumy poskytujú menej informácií o jednotnej agentnosti a o zmyslovom ukotvení ako požiadavkách na vedomie.

[10] Na konferencii NeurIPS som povedal „schopnosti na úrovni rýb“. Zmenil som to na „schopnosti na úrovni myší“ (v zásade asi ťažšia úloha), čiastočne preto, že viac ľudí je presvedčených o tom, že myši sú vedomé, ako o tom, že ryby sú vedomé, a čiastočne preto, že na problematike kognitívnych schopností myší sa pracuje oveľa viac ako na kognitívnych schopnostiach rýb.

Zhrniem jednotlivé výzvy, pričom po štyroch základných výzvach nasleduje sedem inžiniersky zameraných výziev a dvanásta výzva vo forme otázky.

1. Dôkazy: Vypracovať referenčné hodnoty pre vedomie.
2. Teória: Vypracovať lepšie vedecké a filozofické teórie vedomia.
3. Interpretovateľnosť: Pochopiť, čo sa deje vo vnútri LLM.
4. Etika: Mali by sme vytvoriť vedomú umelú inteligenciu?

5. Vybudovať bohaté modely vnímania, jazyka a akcie vo virtuálnych svetoch.
6. Vybudovať LLM+ s robustnými modelmi sveta a seba-modelmi.
7. Vybudovať LLM+ so skutočnou pamäťou a skutočnou rekurenciou.
8. Vybudovať LLM+ s globálnym pracovným priestorom.
9. Vybudovať LLM+, ktoré sú zjednotenými modelmi agentov.
10. Vybudovať LLM+, ktoré opisujú netrénované vlastnosti vedomia.
11. Vybudovať LLM+ so schopnosťami na úrovni myši.

12. Ak to nestačí pre vedomú umelú inteligenciu: čo ešte chýba?

K dvanástej výzve: predpokladajme, že v nasledujúcom desaťročí alebo dvoch splníme všetky technické výzvy v jednom systéme. Budeme mať potom vedomé systémy umelej inteligencie? Nie všetci budú súhlasiť s tým, že áno. Ale ak niekto nesúhlasí, môžeme sa opäť spýtať: Čo je ono X, ktoré chýba? A dalo by sa toto X zabudovať do systému umelej inteligencie?

Podľa môjho názoru, aj keď v najbližšom desaťročí nebudeme mať umelú všeobecnú inteligenciu na úrovni človeka, môžeme mať systémy, ktoré sú serióznymi kandidátmi na vedomie. Na ceste k vedomiu v systémoch strojového učenia je množstvo výziev, ale ich splnenie prináša možný výskumný program smerujúci k vedomej umelej inteligencii.

Na záver zopakujem etickú výzvu.[11] Netvrdím, že by sme mali pokračovať v tomto výskumnom programe. Ak si myslíte, že vedomá umelá inteligencia je žiadúca, tento program môže slúžiť ako akýsi orientačný plán, ako sa k nej dostať. Ak si myslíte, že vedomá umelá inteligencia je niečo, čomu sa treba vyhnúť, potom program môže poukázať na cesty, ktorým je lepšie sa vyhnúť. Pri vytváraní modelov agentov by som bol obzvlášť opatrný. Myslím si, že je pravdepodobné, že výskumníci sa budú venovať mnohým prvkom tohto výskumného programu bez ohľadu na to, či to považujú za snahu o dosiahnutie vedomia umelej inteligencie. Mohlo by to byť katastrofou, ak by sa na vedomie AI narazilo nevedomky a nereflektovane. Dúfam teda, že explicitné vyjadrenie

[11] Tento posledný odsek je dodatkom k tomu, čo som prezentoval na konferencii NeurIPS.

týchto možných ciest nám aspoň pomôže uvažovať o vedomej umelej inteligencii kriticky a pristupovať k týmto otázkam opatrne.

Doslov

Ako sa veci majú teraz, v júli 2023, osem mesiacov po mojej prednáške na konferencii NeurIPS z konca novembra 2022? Hoci nové systémy, ako napríklad GPT-4, majú stále množstvo nedostatkov, predstavujú významný pokrok v niektorých oblastiach, o ktorých sa hovorí v tomto článku. Určite vykazujú sofistikovanejšie konverzačné schopnosti. Ak som povedal, že výkon GPT-3 sa často zdal byť na úrovni šikovného dieťaťa, výkon GPT-4 sa často (nie vždy) zdá byť na úrovni vzdelaného mladého dospelého človeka. Pokrok nastal aj v multimodálnom spracovaní a v modelovaní agentov, a v menšej miere aj v ostatných dimenziách, o ktorých som hovoril. Nemyslím si, že tieto pokroky nejako zásadne menia moju analýzu, ale keďže pokrok bol rýchlejší, ako sa očakávalo, je rozumné skrátiť očakávané časové horizonty. Ak je to tak, moje predpovede na konci tohto článku by mohli dokonca vyznieť trochu konzervatívne.

Preložila Eva Dědečková

Literatúra

ABDOU, M. – KULMIZEV, A. et al. (2021): Can Language Models Encode Perceptual Structure without Grounding? A Case Study in Color. In: *Proceedings of the 25th Conference on Computational Natural Language Learning, November 10 – 11, 2021.* [Online.] Association for Computational Linguistics, 109 – 132. DOI: https://doi.org/10.48550/arXiv.2109.06129

BAARS, B. J. (1988): *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

BENDER, E. M. – GEBRU, T. et al. (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 610 – 623. DOI: https://doi.org/10.1145/3442188.3445922

BENDER, E. M. – KOLLER, A. (2020): Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5 – 10, 2020.* [Online.] Association for Computational Linguistics, 5185 – 5198. DOI: https://doi.org/10.18653/v1/2020.acl-main.463

BENGIO, Y. (2017): The Consciousness Prior. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.1709.08568

BERKOWITZ, R. (2022): How to talk with an AI: A Deep Dive Into "Is LaMDA Sentient?". [Online.] *Medium*. Available at: https://medium.com/curiouserinstitute/guide-to-is-lamda-sentient-a8eb32568531

BIRCH, J. (2021): The Search for Invertebrate Consciousness. *Nous*, 56, 133 – 153. DOI: https://doi.org/10.1111/nous.12351

BLOCK, N. (2009): Comparing the Major Theories of Consciousness. In: Gazzaniga, M. (ed.): *The Cognitive Neurosciences IV*. Cambridge, Massachusetts: MIT Press. DOI: https://doi.org/10.7551/mitpress/8029.003.0099

BOURGET, D. – CHALMERS, D. (2023): Philosophers on Philosophy: The 2020 PhilPapers Survey. *Philosophers' Imprint*, 23 (11), 1 – 53. DOI: https://doi.org/10.3998/phimp.2109

BROWN, T. et al. (2020): Language Models Are Few Shot Learners. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2005.14165

BROWNING, J. – LECUN, Y. (2022): AI and the Limits of Language. [Online.] *Noēma*. Available at: https://www.noemamag.com/ai-and-the-limits-of-language/

DAHAENE, S. (2014): *Consciousness and the Brain.* New York: Penguin.

FRANKEL, J. C. et al. (2022): An Academic Survey on Theoretical Foundations, Common Assumptions and the Current State of Consciousness Science. *Neuroscience of Consciousness*, 1, 1 – 13. DOI: https://doi.org/10.1093/nc/niac011

FRANKISH, K. (ed.) (2017): *Illusionism as a Theory of Consciousness*. Exeter: Imprint Academic.

GENNARO, R. (ed.) (2004): *Higher-Order Theories of Consciousness: An Anthology*. Amsterdam: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/aicr.56

GOYAL, A. – DIDOLKAR, A. et al. (2021): Coordination among Neural Modules through a Shared Global Workspace. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2103.01197

GRAZIANO, M. (2019): *Rethinking Consciousness*. New York: W.W. Norton.

HARNAD, S. (1990): The Symbol-Grounding Problem. *Physica D: Nonlinear Phenomena*, 42 (1 – 3), 335 – 346. DOI: https://doi.org/10.1016/0167-2789(90)90087-6

HOCHREITER, S. – SCHMIDHUBER, J. (1997): Long Short-Term Memory. *Neural Computation*, 9 (8), 1735 – 1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

CHALMERS, D. J. (2020): GPT-3 and General Intelligence. [Online.] *Daily Nous*, July 30, 2020. Available at: https://dailynous.com/2020/07/30/philosophers-gpt-3/

CHALMERS, D. J. (2022): *Reality+: Virtual Worlds and the Problems of Philosophy*. New York: W.W. Norton.

CHALMERS, D. J. (2023): Does Thought Require Sensory Grounding: From Pure Thinkers to Large Language Models. In: *Proceedings and Addresses of the American Philosophical Association, vol. 97*. [Online.] American Philosophical Association, 22 – 45. DOI: https://doi.org/10.48550/arXiv.2408.09605

JAEGLE et al. (2021): Perceiver IO: A General Architecture for Structured Inputs and Outputs. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2107.14795

JULIANI, A. et al. (2021): The Perceiver Architecture is a Functional Global Workspace. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44, 955 – 961. [Online.] Available at: https://escholarship.org/uc/item/2g55b9xx

KAGAN, B. et al. (2022): In Vitro Neurons Learn and Exhibit Sentience When Embodied in a Simulated Game-World. *Neuron*, 110 (23), 3952 – 3969. DOI: https://doi.org/10.1016/j.neuron.2022.09.001

LAKE, B. – MURPHY, G. (2021): Word Meaning in Minds and Machines. *Psychological Review*, 130 (2), 401 – 431. DOI: https://psycnet.apa.org/doi/10.1037/rev0000297

LAMME, V. A. (2010): How Neuroscience Will Change Our View on Consciousness. *Cognitive Neuroscience*, 1 (3), 204 – 220. DOI: https://doi.org/10.1080/17588921003731586

LEMOINE, B. (2022): Is LaMDA Sentient? An Interview. [Online.] *Medium*. Available at: https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

LI, B. et al. (2021): Implicit Representations of Meaning in Neural Language Models. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1813 – 1827. [Online.] Association for Computational Linguistics. Available at: https://aclanthology.org/2021.acl-long.143/

LI, K. – HOPKINS, A. K. et al. (2022): Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2210.13382

LIAO, M. (2020): *Ethics of Artificial Intelligence*. Oxford: Oxford University Press.

LONG, R – BUTLIN, P. et al. (2023): Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2308.08708

LYCAN, W. (2023): Representational Theories of Consciousness. In: Zalta, E. N. – Nodelman, U. (eds.): *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*. Available at: https://plato.stanford.edu/archives/win2023/entries/consciousness-representational/

NAGEL, T. (1974): What Is It Like to Be a Bat? *Philosophical Review*, 83 (4), 435 – 450. DOI: https://doi.org/10.2307/2183914

PATEL, R. – PAVLICK, E. (2022): Mapping Language Models to Grounded Conceptual Spaces. [Poster.] *International Conference on Learning Representations*, 1 – 21. Available at: https://openreview.net/forum?id=gJcEM8sxHK

REED, S. et al. (2022): "A Generalist Agent. Transactions on Machine Learning Research." [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2205.06175

SCHNEIDER, S. (2019): *Artificial You*. Princeton, New Jersey: Princeton University Press.

TONONI, G. (2004): An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5, article 42, 1 – 22. DOI: https://doi.org/10.1186/1471-2202-5-42

TURING, A. (1950): Computing Machinery and Intelligence. *Mind*, 59 (236), 433 – 460. DOI: https://doi.org/10.1093/mind/LIX.236.433

ZADOR, A. et al. (2022): Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI revolution. [Online.] *arXiv*. DOI: https://doi.org/10.48550/arXiv.2210.08340

Pôvodne publikované v *Boston Review*, 9. augusta 2023; toto je preklad editovanej verzie prednášky prezentovanej na konferencii o Neural Information Processing Systems (NeurIPS), 28. novembra 2022, s drobnými doplneniami a úpravami.

David Chalmers
Department of Philosophy
New York University
5 Washington Place
New York, NY 10003
USA
e-mail: chalmers@nyu.edu
ORCID ID: https://orcid.org/0000-0003-4190-3161